SIS | 2022

51st Scientific Meeting
of the Italian Statistical Society

Caserta, 22-24 June

Università degli Studi della Campania Luigi Vanvitelli

SIS Società Italiana di Statistica

www.unicampania.it

Book of the Short Papers

Editors: Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde

Città di Caserta

800 ANNI 1222-2022 Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Stata

sas

Università degli Studi del Sannio Benevento

Pearson

Matilde Bini (Chair of the Program Committee) - *Università Europea di Roma*
Rosanna Verde (Chair of the Local Organizing Committee) - *Università della Campania "Luigi Vanvitelli"*

PROGRAM COMMITTEE

Matilde Bini (Chair), Giovanna Boccuzzo, Antonio Canale, Maurizio Carpita, Carlo Cavicchia, Claudio Conversano, Fabio Crescenzi, Domenico De Stefano, Lara Fontanella, Ornella Giambalvo, Gabriella Grassia - Università degli Studi di Napoli Federico II, Tiziana Laureti, Caterina Liberati, Lucio Masserini, Cira Perna, Pier Francesco Perri, Elena Pirani, Gennaro Punzo, Emanuele Raffinetti, Matteo Ruggiero, Salvatore Strozza, Rosanna Verde, Donatella Vicari.

LOCAL ORGANIZING COMMITTEE

Rosanna Verde (Chair), Antonio Balzanella, Ida Camminatiello, Lelio Campanile, Stefania Capecchi, Andrea Diana, Michele Gallo, Giuseppe Giordano, Ferdinando Grillo, Mauro Iacono, Antonio Irpino, Rosaria Lombardo, Michele Mastroianni, Fabrizio Maturo, Fiammetta Marulli, Paolo Mazzocchi, Marco Menale, Giuseppe Pandolfi, Antonella Rocca, Elvira Romano, Biagio Simonetti.

ORGANIZERS OF SPECIALIZED, SOLICITED, AND GUEST SESSIONS

Arianna Agosto, Raffaele Argiento, Massimo Aria, Rossella Berni, Rosalia Castellano, Marta Catalano, Paola Cerchiello, Francesco Maria Chelli, Enrico Ciavolino, Pier Luigi Conti, Lisa Crosato, Marusca De Castris, Giovanni De Luca, Enrico Di Bella, Daniele Durante, Maria Rosaria Ferrante, Francesca Fortuna, Giuseppe Gabrielli, Stefania Galimberti, Francesca Giambona, Francesca Greselin, Elena Grimaccia, Raffaele Guetto, Rosalba Ignaccolo, Giovanna Jona Lasinio, Eugenio Lippiello, Rosaria Lombardo, Marica Manisera, Daniela Marella, Michelangelo Misuraca, Alessia Naccarato, Alessio Pollice, Giancarlo Ragozini, Giuseppe Luca Romagnoli, Alessandra Righi, Cecilia Tomassini, Arjuna Tuzzi, Simone Vantini, Agnese Vitali, Giorgia Zaccaria.

ADDITIONAL COLLABORATORS TO THE REVIEWING ACTIVITIES

Ilaria Lucrezia Amerise, Ilaria Benedetti, Andrea Bucci, Annalisa Busetta, Francesca Condino, Anthony Cossari, Paolo Carmelo Cozzucoli, Simone Di Zio, Paolo Giudici, Antonio Irpino, Fabrizio Maturo, Elvira Romano, Annalina Sarra, Alessandro Spelta, Manuela Stranges, Pasquale Valentini, Giorgia Zaccaria.

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891932310

Contents

Preface **XXVII**

1 Plenary Sessions 1

Causal inference in air pollution epidemiology 2
Francesca Dominici

Clustering of Attribute Data and Network 11
Anuška Ferligoj

Bayesian approaches for capturing the heterogeneity of neuroimaging experiments 17
Francesco Denti, Laura D'Angelo and Michele Guindani

2 Specialized Sessions 30

Advances in Bayesian nonparametric methodology 31

Repulsive mixture models for high-dimensional data 32
Lorenzo Ghilotti, Mario Beraha and Alessandra Guglielmi

Bayesian nonparametric mixtures of directed acyclic graph models 37
Federico Castelletti and Guido Consonni

Bayesian Clustering of Brain Regions via Extended Stochastic Block Models 45
Sirio Legramanti, Tommaso Rigon and Daniele Durante

Data Science skills for next generation statisticians 52

Cluster based oversampling for imbalanced learning 53
Gioia Di Credico and Nicola Torellii

Estimating the effect of remote teaching for university students through generalised linear mixed models 65
Silvia Bacci, Bruno Bertaccini, Simone Del Sarto, Leonardo Grilli and Carla Rampichini

Perceived stress across EU countries: does working from home impact? 71
Stefania Capecchi, Francesca Di Iorio and Nunzia Nappo

Investigating effects of air pollution on health: a challenge for statisticians 77

Investigating effect of air pollution on health via Spatial-Resolution Varying Coefficient Models 78
Garritt L. Page and Massimo Ventrucci

A statistical framework for evaluating health effect of PM sources 84
Monica Pirani, Georges Bucyibaruta, Gary Fuller, David Green, Anja Tremper, Christina Mitsakou and Marta Blangiardo

Adjusting for unmeasured spatial confounding through shrinkage methods 91
Pasquale Valentini, Alexandra M. Schmidt, Carlo Zaccardi and Luigi Ippoliti

Explainable Artificial Intelligence methods 98

Multidimensional Time Series Analysis via Bayesian Matrix Auto Regression 99
Alessandro Celani and Paolo Pagnottoni

Advances in Classification and Data Analysis 109

Optimizing time slots in scientific meetings: a Latent Dirichlet allocation approach 110
Luca Frigau

Clustering artists based on the energy distributions of their songs on Spotify via the Common Atoms Model 121
Francesco Denti, Federico Camerlenghi, Michele Guindani and Antonietta Mira

Hidden markov models for four-way data 127
Salvatore D. Tomarchio, Antonio Punzo and Antonello Maruotti

Family demography 133

Does family of origin make the difference in occupational outcomes? 134
Annalisa Busetta, Elena Fabrizi, Isabella Sulis and Giancarlo Ragozini

Is there a cultural driver pushing Italian low fertility? 144
Francesca Luppi, Alessandro Rosina and Maria Rita Testa

Unpaid family work and the subjective well-being of Italian women during lockdown 155
Marina Zannella, Erica Aloé, Marcella Corsi and Alessandra de Rose

New Frontiers in the theory of composite indicators 164

Methodological PLS-PM Framework for Model Based Composite Indicators 165
Rosanna Cataldo

Open issues in composite indicators construction 176
Leonardo Salvatore Alaimo

The posetic approach to the construction of socio-economic indicators: open issues and research opportunities 186
Marco Fattore

Advances in complex sampling strategies 197

Random forest model-assisted estimation for finite population totals 198
Mehdi Dagdoug, Camelia Goga and David Haziza

Design-based consistency of the Horvitz-Thompson estimator in spatial sampling 208
Lorenzo Fattorini

The responsive-adaptive survey design approach for planning the permanent census of population and housing 216
Claudia De Vitiis, Stefano Falorsi, Alessio Guandalini, Francesca Inglese, Paolo Righi and Marco D. Terribili

Socio-demographic aspects of aging in Italy 228

Socio-economic and spatial stratification of frailty in the older population 229
Margherita Silan

Time allocation and wellbeing in later life: the case of Italy 241
Annalisa Donno and Maria Letizia Tanturri

The role played by migration and fertility on Italy's demographic aging trends: a provincial-level analysis 250
Thaís García-Pereiro and Anna Paterno

New challenges in the labour market 260

Detecting changes and evolution in specialized professional figures: an application on the Italian IT & Digital sector 261
Andrea Marletta

How did the COVID-19 pandemic affect the genderpay gap in EU countries? 272
Antonella Rocca, Paolo Mazzocchi, Giovanni De Luca, Rosalia Castellano and Claudio Quintano

Skill Similarities and Dissimilarities in Online Job Vacancy Data across Italian Regions 284
Adham Kahlawi, Lucia Buzzigoli, Laura Grassini and Cristina Martelli

Small area estimation methods with socioeconomic applications 292

Exploring Small Area Estimation techniques to address uncertainty in Spatial Price Indexes 293
Ilaria Benedetti and Federico Crescenzi

Small Area Estimation of Relative Inequality Indices using Mixture of Beta 301
Silvia De Nicolò and Silvia Pacei

Inference for big data assisted by small area methods: an application to OBEC (on-line based enterprise characteristics) 305
Monica Pratesi, Francesco Schirripa Spagnolo, Gaia Bertarelli, Stefano Marchetti, Monica Scannapieco, Nicola Salvati and Donato Summa

Statistical methods and models for Sports Analytics 312

The 'hot shoe' in soccer penalty shootouts 313
Andreas Groll and Marius Otting

G-RAPM: revisiting player contributions in regularized adjusted plus-minus models forbasketball analytics 319
Luca Grassetti

Formative vs Reflective constructs: a CTA-PLS approach on a goalkeepers' performance model 323
Mattia Cefis and Eugenio Brentari

Integrating available Data Sources for Official Statistics 329

The Use of Administrative Data for the Estimation of Italian Usually Resident Population 330
Marco Caputi, Giampaolo De Matteis, Gerardo Gallo and Donatella Zindato

New frontiers for the analysis of the territorial economic phenomena 339

An empirical tool to classify industries by regional concentration and spatial polarization 340
Diego Giuliani, Maria Michela Dickson, Flavio Santi and Giuseppe Espa

Comparing non-compensatory composite indicators: a case study based on SDG for Mediterranean countries 346
Francesca Mariani, Mariateresa Ciommi, Maria Cristina Recchioni, Giuseppe Ricciardo Lamonica and Francesco Maria Chelli

Evaluating the determinants of innovation from a spatio-temporal perspective. The GWPR approach 354
Gaetano Musella, Giorgia Rivieccio and Emma Bruno

Dimension Reduction for complex data 366

Discrimination and clustering via principal components 367
Nikolay Trendafilov and Violetta Simonacci

Exploratory graph analysis for configural invariance assessment 373
Sara Fontanella, Alex Cucco and Nicola Pronello

Penalized likelihood factor analysis 379
Kei Hirose

3 Solicited Sessions 385

Bayesian nonparametric modelling and learning 386

A regularized-entropy estimator to enhance cluster interpretability in Bayesian nonparametrics 387
Beatrice Franzolini and Giovanni Rebaudo

Exact confidence sets from credible sets with finite amounts of data 399
Bas J. K. Kleijn

Empirical Bayesian analysis of componentwise maxima in multivariate samples 411
Simone A. Padoan and Stefano Rizzelli

Processing of textual data in large corpora 420

Predictive performance comparisons of different feature extraction methods in a financial column corpus 421
Andrea Sciandra and Riccardo Ferretti

Topics and trends in the End-of-Year addresses of the Presidents of the Italian Republic (1949-2021) 428
Matilde Trevisani and Arjuna Tuzzi

Thematic analysis on online education issues during COVID-19 437
Valerio Basile, Michelangelo Misuraca and Maria Spano

What do we learn by applying multiple methods in topic detection? A comparative analysis on a large online dataset about mobility electrification 446
Fabrizio Alboni, Margherita Russo and Pasquale Pavone

Businesses in industry: new challenges in sustainability, innovation, performance and competitiveness 454

Multidimensional assessment of Eco-Innovation and its link with Marketing Innovations 455
Ida D'Attoma and Marco Ieva

Circular Economy practices in the European SMEs: company-level and country-level drivers 462
Francesca Bassi, Josè G. Dias and Nunzio Tritto

The employment effects of Italian Jobs Act. An ex-post impact evaluation 474
Alessandro Zeli and Leopoldo Nascia

Statistics for finance: new models, new data 482

The News-Jumps Relationship in the Cryptocurrency Market 483
Ahmet Faruk Aysan, Massimiliano Caporin, Oguzhan Cepni, and Francesco Poli

A weighted quantile approach to Expected Shortfall forecasting 489
Giuseppe Storti and Chao Wang

Smooth and abrupt dynamics in financial volatility: the MS-MEM-MIDAS 492
Giampiero M. Gallo, Edoardo Otranto and Luca Scaffidi Domianello

The tail index and related quantities for volatility models 501
Fabrizio Laurini

Bayesian inference for complex random structures 507

Bayesian nonparametric modeling of mortality curves via functional Dirichlet processes 508
Emanuele Aliverti and Bruno Scarpa

Bayesian nonparametric clustering of spatially-referenced spike train data 514
Laura D'Angelo

Bayesian Analysis of Mortality in Iceland via Locally Adaptive Splines 520
Federico Pavone and Sirio Legramanti

Advances in clustering 526

A Two-step Latent Class Approach with Measurement Equivalence Testing 527
Zsuzsa Bakk, Roberto Di Mari, Jennifer Oser and Marc Hooghe

Group-wise penalized estimation schemes in model-based clustering 534
Alessandro Casa, Andrea Cappozzo and Michael Fop

Extending finite mixtures of latent trait analyzers for bipartite networks 540
Dalila Failli, Maria Francesca Marino and Francesca Martella

A Fast Majorization-Minimization Algorithm for Convex Clustering 551
Daniel J.W. Touw, Patrick J.F. Groenen and Yoshikazu Terada

Statistical Methods for Complex Evolutionary Data 558

A FANOVA model with repeated measures for detecting patterns in biomechanical data 559
Ana M. Aguilera, Christian Acal and Manuel Escabias

Modes of variation for Lorenz curves 565
Enea G. Bongiorno and Aldo Goia

Analyzing textual data through Word Embedding: experiences in Istat 571
Mauro Bruno, Elena Catanese, Massimo De Cubellis, Fabrizio De Fausti, Francesco Pugliese, Monica Scannapieco and Luca Valentino

Functional Horvitz-Thompson estimator for convex curves 584
Adelia Evangelista, Francesca Fortuna, Stefano Antonio Gattone and Tonio Di Battista

Children, parents, grandparents: a look on changing relationships 590

Changes in social relationships of Italian older people. Evidence from FSS and SHARE Corona surveys 591
Elvira Pelle, Giulia Rivellini and Susanna Zaccarin

Internet use and contacts with children among older Europeans 600
Bruno Arpino

A time-based comparative approach to study the changing demography of grandparenthood in Italy 607
****Elisa Cisotto, Eleonora Meli and Giulia Cavrini*

Carry that weight: Parental separation and children's Body Mass Index from childhood to young adulthood 616
Marco Tosi

Living conditions, well-being and poverty 622

Analyzing the impact of COVID-19 pandemic on elderly population well-being 623
Gloria Polinesi, Mariateresa Ciommi and Chiara Gigliarano

Exploring sustainable food purchasing behaviour using Italian scanner data 629
Ilaria Benedetti, Alessandro Brunetti, Federico Crescenzi and Luigi Palumbo

The evaluation of heat vulnerability in Friuli-Venezia Giulia 635
Laura Pagani, Maria Chiara Zanarotti and Anja Habus

Data Science for Functional and Complex Data 641

A parsimonious approach to representing functional 642
Enea G. Bongiorno and Aldo Goia

Mixed-effects high-dimensional multivariate regression via group-lasso regularization 648
Francesca Ieva, Andrea Cappozzo, and Giovanni Fiorito

The integration of immigrants in Italy: a multidimensional perspective 654

Albanian, Romanian and Italian women's fertility intentions: a comparative perspective among migrants, stayers and natives 655
Thaís García-Pereiro and Anna Paterno

Does self-employment in the origin-country affect self-employment after migration? Evidence from Italy and Spain 662
Floriane Bolazzi and Ivana Fellini

The impact of integration on immigrants' health behaviours in Italy 675
Giovanni Minchio, Raffaella Rusciani and Teresa Spadea

Migration, gender, and the distribution of paid and unpaid labour. Preliminary perspectives on foreign couples in Italy 687
Rocco Molinari, Agnese Vitali and Ester Gallo

Sampling techniques for big data analysis 695

Non-probability samples and big data: how to use them? 696
Pier Luigi Conti

Combining Big Data with probability survey data: a comparison of methodologies for estimation from non-probability surveys 707
Maria del Mar Rueda, Ramn Ferri-Garcıa and Luis Castro-Martìn

A Bayesian approach for combining probability and non-probability samples surveys 717
Camilla Salvatore, Silvia Biffignandi, Joseph Sakshaug, Bella Struminskaya and Arkadiusz Wisniowski

Big data and Official Statistics: some evidences 723
Paolo Righi, Natalia Golini and Gianpiero Bianchi

The analysis of students performance and behaviour based on large databases 735

Students enrolled in STEM discipline in Italy: patterns of retention, dropout and switch 736
Valentina Tocchioni, Carla Galluccio, Maria Francesca Morabito and Alessandra Petrucci

The routes of Southern Italy University students: an explorative analysis 747
Gabriele Ruiu and Vincenzo Giuseppe Genova

A new bipartite matching approach for record linkage: the case of two big Italian databases 754
Martina Vittorietti, Andrea Priulla, Vincenzo Giuseppe Genova, Giovanni Boscaino and Ornella Giambalvo

Statistical Methods for Science Mapping 761

A word embedding strategy to study the thematic evolution of ageing and healthcare expenditure growth literature 762
Milena Lopreite, Michelangelo Misuraca and Michelangelo Puliga

An automatic approach for bibliographical co-words networks labelling 773
Manuel J. Cobo and Maria Spano

Characterising research areas in the field of AI 780
Alessandra Belfiore, Angelo Salatino and Francesco Osborne

Mapping evolutionary paths of a society: the longitudinal analysis of the Italian Economia Aziendale 786
Corrado Cuccurullo, Luca D'Aniello and Michele Pizzo

Modelling complex structures in ecological data 793

New insights on the ecology and conservation of Mediterranean sharks through the development of Citizen Science networks and new modeling approaches 794
Stefano Moro, Francesco Ferretti, Francesco Colloca

An overdispersed Poisson model for forest fires occurrences in Southern Italian municipalities 798
Crescenza Calculli and Serena Arima

Assessment of the impact of anthropic pressures on the Giglio island meadow of Posidonia oceanica 804
Gianluca Mastrantonio, Daniele Ventura, Gianluca Mancini and Giandomenico Ardizzone

Accounting for observation processes in spatio-temporal ecological data 811
Janine Illian

Statistics and indicators for the recovery and resilience plan 815

The prominence of statistical information for the monitoring and effective implementation of the NRRP 816
Andrea Petrella

Big Data Analytics in mobile cellular networks as enabler for innovative statistics to evaluate the effects of Recovery and Resilience Plan actions 819
Andrea Zaramella, Dario Di Sorte, Denis Cappellari and Bruno Zamengo

Measuring the digital transition within the PA: proposals comparison 823
Susanna Traversa and Enrico Ivaldi

Guest Session - European Network for Business and Industrial Statistics (ENBIS) 828

Interpretability in functional clustering with an application to resistance spot welding process in the automotive industry 829
Christian Capezza, Fabio Centofanti, Antonio Lepore and Biagio Palumbo

Statistical process monitoring of thermal images in additive manufacturing: a nonparametric solution for in-situ monitoring 835
Panagiotis Tsiamyrtzis, Marco Luigi Giuseppe Grasso and Bianca Maria Colosimo

Guest Session - International Biometric Society (IBS) - Italian region 837

Multiple arrows in the Bayesian quiver: Bayesian learning of partially directed structures from heterogeneous data 838
Luca La Rocca, Federico Castelletti, Stefano Peluso, Francesco Claudio Stingo and Guido Consonni

4 Contributed Sessions 844

Applications in Machine Learning 845

A neural network approach to survival analysis with time-dependent covariates for modelling time to cardiovascular diseases in HIV patients 846
Federica Corso, Agostino Lurani Cernuschi, Laura Galli, Chiara Masci, Camilla Muccini, Anna Maria Paganoni and Francesca Ieva

Analyzing the Correlation Structure of Financial Markets Using a Quantile Graphical Model 852
Beatrice Foroni, Luca Merlo and Lea Petrella

Neural Network for statistical process control of a multiple stream binomial process with an application to HVAC systems in passenger rail vehicles 858
Gianluca Sposito, Antonio Lepore, Biagio Palumbo and Giuseppe Giannini

Sparse signal extraction via variational SVM 864
Cristian Castiglione and Mauro Bernardi

Bayesian modelling and inference 1 870

Bayesian Inference for the Multinomial Probit Model under Gaussian Prior Distribution 871
Augusto Fasano, Giovanni Rebaudo and Niccolo Anceschi

Mapping Indicators on the Unit Interval: the tipsae Shiny App 877
Silvia De Nicolò and Aldo Gardini

A Bayesian spatio–temporal model of PM10 pollutant in the Po Valley 883
Matteo Gianella, Alessandra Guglielmi and Giovanni Lonati

Construction if a proper prior for a Bayesian envelope model 889
Andrea Mascaretti

Hilbert principal component regression for bimodal bounded responses 895
Enea G. Bongiorno, Agnese M. Di Brisco, Aldo Goia, and Sonia Migliorati

Methods of causal inference 901

Bayesian causal mediation analysis through linear mixed-effect models 902
Chiara Di Maria, Antonino Abbruzzo and Gianfranco Lovison

Bootstrap-aggregated adjustment set selection 908
Lorenzo Giammei

Exploiting partial knowledge to evaluate the average causal effect via an ABC perspective 914
Giulia Cereda, Fabio Corradi and Cecilia Viscardi

Intertemporal propensity score matching for casual inference: an application to covid-19 lockdowns and air pollution in Northern Italy 920
Daniele Bondonio and Paolo Chirico

Methods for Spatio-temporal data 926

Local Spatio-Temporal Log-Gaussian Cox Processes for seismic data analysis 927
Nicoletta D'Angelo, Giada Adelfio, and Jorge Mateu

Spatial explorative analysis of thyroid cancer in Sicilian volcanic areas 933
Francesca Bitonti and Angelo Mazza

Using geo-spatial topic modelling to understand the public view of Italian Twitter users: a climate change application 939
Yuri Calleo and Francesco Pilla

Comparing local structures of spatio-temporal point processes on linear networks 945
Nicoletta D'Angelo, Giada Adelfio, and Jorge Mateu

DISTATIS-based spatio-temporal clustering approach: an application to business cycles' time series 951
Raffaele Mattera and Germana Scepi

Developments in composite indicators 957

Bayesian Networks for monitoring the gender gap 958
Flaminia Musella, Lorenzo Giammei, Silvana Romio, Fulvia Mecatti and Paola Vicard

An Alternative Aggregation Function for the UNDP Human Development Index 964
Manuela Scioni and Paola Annoni

An ultrametric model for building a composite indicator system to study climate change in European countries 970
Giorgia Zaccaria and Pasquale Sarnacchiaro

Functional Weighted Malmquist Productive Index: a proposal for a dynamic composite indicator 975
Annalina Sarra, Eugenia Nissi and Tonio Di Battista

CFA & PLS-PM for UX-AI Product infused 981
Emma Zavarrone and Rosanna Cataldo

Fertility, adulthood, and economic uncertainty 987

Uncertainty and fertility intentions: a comparison between the Great Recession and the Covid-19 crisis 988
Chiara Ludovica Comolli

Interpreting the relationship between life course trajectories and explanatory factors. An example on the transition to adulthood 996
Danilo Bolano, Matthias Studer and Reto Buergin

The relationship between economic news and fertility: the case of Germany 1002
Maria Francesca Morabito, Raffaele Guetto, Matthias Vollbracht and Daniele Vignoli

Leaving home among Millennials in Italy: does economic uncertainty matter? 1008
Silvia Meggiolaro and Fausta Ongaro

Adverse pregnancy outcomes in The United Kingdom following unexpected job loss 1014
Alessandro Di Nallo and Selin Koksal

Bayesian modelling and inference 2 1020

A Bayesian beta linear model to analyze fuzzy rating responses 1021
Antonio Calcagnì, Massimiliano Pastore, Gianmarco Altoe and Livio Finos

A Mixture Model for Multi-Source Cyber-Vulnerability Assessment 1028
Mario Angelelli, Serena Arima and Christian Catalano

Hierarchical Bayesian models for analysing fish biomass data 1034
Rita Fici, Antonino Abbruzzo, Luigi Augugliaro and Giacomo Milisenda

Insights into the derivative-based method for nonlinear mediation models 1040
Claudio Rubino and Chiara Di Maria

An exploration of Approximate Bayesian Computation (ABC) and dissimilarities 1046
Laura Bondi, Marco Bonetti and Raffaella Piccarreta

Advances in Categorical and Preference data 1052

On the predictability of a class of ordinal data models 1053
Rosaria Simone and Domenico Piccolo

Multivariate analysis of binary ordinal data using graphical models 1059
Camilla Caroni, Fabio Alberto Comazzi, Andrea Deretti and Federico Castelletti

Multinomial Thompson Sampling for adaptive experiments with rating scales 1065
Nina Deliu

Ranking extraction in nested partially ordered data systems 1071
Marco Fattore, Barbara Cavalletti, Matteo Corsi and Alessandro Avellone

Towards the definition of distance measures in the preference-approval structures 1077
Alessandro Albano, Mariangela Sciandra and Antonella Plaia

Covid-19 Assessment and Evaluation 1 1083

Covid-19 impact assessment and inequality decomposition methods 1084
Federico Attili and Michele Costa

Multiversal methods for model selection: COVID-19 vaccine coverage and relative risk reduction 1090
Venera Tomaselli and Giulio Giacomo Cantone

Efficiency and feasibility of two stage sampling designs for estimating SARS-CoV-2 epidemic 1096
Pietro Demetrio Falorsi, Vincenzo Nardelli and Giuseppe Arbia

Evaluating the impacts of Covid-19 on the overall Italian death process via Functional Data Analysis 1102
Riccardo Scimone, Alessandra Menafoglio, Laura M. Sangalli and Piercesare Secchi

Developing countries, migration and migrants 1107

Domestic violence in Africa: a glance through the DHS survey 1108
Micaela Arcaio, Daria Mendola and Anna Maria Parroco

Inequalities in undernutrition among Roma and non-Roma children in Western Balkans: an analysis of the determinants 1114
Annalisa Busetta, Valeria Cetorelli and Chiara Puglisi

The manual, communicative and quantitative abilities of native and foreign workers according to their level of education in Italy 1120
Camilla Pangallo, Oliviero Casacchia and Corrado Polli

HIV Prevalence in some African Territories: Socio-Economic Drivers 1126
Micaela Arcaio, Daria Mendola and Anna Maria Parroco

A longitudinal cross country comparison of migrant integration policies via Mixture of Matrix-Normals 1132
Leonardo Salvatore Alaimo, Francesco Amato and Emiliano Seri

Education and job placement 1138

Measuring happiness at work with categorical Principal Component Analysis 1139
Ulpiana Kocollari, Maddalena Cavicchioli and Fabio Demaria

Early and accurate: a Machine Learning approach to predict students' final outcome with registry data 1146
Lidia Rossi, Marta Cannistrà and Tommaso Agasisti

Students' experience with distance learning during Covid 19 pandemic in Southern Italy 1153
Angela Maria D'Uggento and Nunziata Ribecco

Time series methods and Applications 1159

Trend and cycle decomposition in nonlinear time series 1160
Maddalena Cavicchioli

Asymptotic properties of the SETAR parameters: a new approach 1166
Marcella Niglio and Guy Mélard

Food prices forecast using post-sampled crowdsourced data with Reg-ARMA model: the case of Nigeria 1172
Ilaria Lucrezia Amerise, Gloria Solano Hermosilla, Vincenzo Nardelli and Giuseppe Arbia

Universal change point testing for dependent data 1178
Federica Spoto, Alessia Caponera and Pierpaolo Brutti

Change point detection in fruit bioimpedance using a three-way panel model 1184
F. Marta L. Di Lascio and Selene Perazzini

Bayesian modelling and inference 3 1190

A dynamic power prior approach to non-inferiority trials for normal means with unknown variance 1191
Francesco Mariani, Fulvio De Santis and Stefania Gubbiotti

Bayesian Change-Point Detection for a Brownian Motion with a Total Miss Criterion 1197
Bruno Buonaguidi

On the comparison of alternative Bayesian measures of posterior discrepancy 1203
Fulvio De Santis and Stefania Gubbiotti

A Bayesian Test for the comparison of two independent populations 1209
Mara Manca, Silvia Columbu and Monica Musio

A contribution to the L. J. Savage problem 1215
Francesco Bertolino, Silvia Columbu and Mara Manca

Methods for Complex Data 1221

Optimization of delayed rejection adaptive metropolis 1222
Daniele Raffo and Antonietta Mira

Dealing with multicollinearity and outliers in multinomial logit model: a simulation study 1228
Ida Camminatiello and Antonio Lucadamo

A tool to validate the assumptions on ratios of nearest neighbors' distances: the Consecutive Ratio Paths 1233
Francesco Denti and Antonietta Mira

Dimensionality reduction and visualization for interval-valued data via midpoints-ranges principal component analysis 1239
Viviana Schisa, Alfonso Iodice D'Enza and Francesco Palumbo

Data-driven design-based mapping of forest resources 1245
Sara Franceschi, Rosa Maria Di Biase, Lorenzo Fattorini, Marzia Marcheselli and Caterina Pisani

Environmental data and Climate change 1252

Ensemble model output statistics for temperature forecasts in Veneto 1253
Gaetan Carlo, Giummole Federica, Mameli Valentina and Siad Si Mokrane

State of the urban Environment in Italy. A comparative analysis of selected composite indicators 1259
Giuseppe Lecardane

A Functional Data Analysis approach for Climate Model Selection: the case study of Campania Region 1266
Veronica Villani, Elvira Romano and Paola Mercogliano

Evolution of scientific literature on climate change: a bibliometric analysis 1273
Gianpaolo Zammarchi, Giulia Contu, Maurizio Romano

Energy and material demand of the Italian Regions 1279
Flora Fullone, Giulia Iorio, Assunta Lisa Carulli

Health and survivorship 1285

Increasing Inequalities in Mortality by Socioeconomic Position in Italy 1286
Chiara Ardito, Nicolás Zengarini, Roberto Leombruni, Angelo d'Errico and Giuseppe Costa

The role of health conditions in the relationship between socio-economic status and well-being: the counterfactual approach in mediation models 1296
Sara Manzella and Margherita Silan

Excess economic burden of multimorbidity: a population-based study in Italy 1302
Chiara Seghieri, Niccolò Borri, Gaia Bertarelli and Sabina Nuti

Depression-free life expectancy among 50 and older Americans by gender, race/ethnicity and education: the effect of marital disruption 1308
Alessandro Feraldi and Cristina Giudici

Disability-free grandparenthood in Italy. Trends and gender differences 1314
Margherita Moretti, Elisa Cisotto and Alessandra De Rose

Advances in regression models 1320

Semiparametric M-quantile regression for modelling georeferenced housing price data 1321
Riccardo Borgoni, Antonella Carcagnì, Alessandra Michelangeli, Nicola Salvati and Francesco Schirripa Spagnolo

Resampling-based inference for high-dimensional regression 1327
Anna Vesel, Jelle J. Goeman, Angela Andreella and Livio Finos

Quantile regression coefficient modeling for counts to evaluate the productivity of university students 1333
Viviana Carcaiso and Leonardo Grilli

Adaptive smoothing spline using non-convex penalties 1339
Daniele Cuntrera and Vito M.R. Muggeo

Conditional tests for generalized linear models 1345
Riccardo De Santis, Jelle J. Goeman, Anna Vesely and Livio Finos

Methods and applications in economics and finance 1351

Mixed models for anomaly detection in anti-money laundering aggregate reports 1352
Stefano Iezzi and Marianna Siino

On the drivers of Greenwashing risk: evidence from Eurostoxx600 1358
Yana Kostiuk, Costanza Bosone and Paola Cerchiello

Modelling Financial Returns with Finite Mixtures of GED 1364
Pierdomenico Duttilo and Stefano Antonio Gattone

Risk Parity strategy for portfolio construction: a kurtosis-based approach 1370
Maria Debora Braga, Consuelo Rubina Nava and Maria Grazia Zoia

Fully reconciled probabilistic GDP forecasts from Income and Expenditure sides 1376
Tommaso Di Fonzo and Daniele Girolimetto

Latent Class models 1382

Latent thresholds model in classification tasks 1383
Giuseppe Mignemi, Andrea Spoto and Antonio Calcagnì

Adaptive filters for time-varying correlation parameters 1389
Michele Lambardi di San Miniato, Ruggero Bellio, Luca Grassetti and Paolo Vidoni

Bayesian structural learning for Latent Class Model with an application to Record Linkage 1395
Davide Di Cecco

Multilevel Latent Class modelling to advise students in self-learning platforms: an application in the context of learning Statistics 1401
Roberto Fabbricatore, Zsuzsa Bakk, Roberto Di Mari, Mark de Rooij and Francesco Palumbo

Latent Markov models with associated mixed responses 1407
Alfonso Russo and Alessio Farcomeni

Methods for health studies 1413

Beyond the fragility index 1414
Piero Quatto and Enrico Ripamonti

Evaluation of the diagnostic-therapeutic paths for schizophrenic patients through state sequences analysis 1419
Laura Savaré, Giovanni Corrao and Francesca Ieva

Optimal timing of bone-marrow transplant in myelodysplastic syndromes through multi-state modeling and microsimulation 1425
Caterina Gregorio, Marta Spreafico and Francesca Ieva

A fully Bayesian approach for sample size determination of Poisson clinical trials 1431
Susanna Gentile and Valeria Sambucini

Compartmental models in epidemiology: Application on Smoking Habits in Tuscany 1437
Alessio Lachi, Cecilia Viscardi, Maria Chiara Malevolti, Giulia Carreras and Michela Baccini

Covid-19 Assessment and Evaluation 2 1443

We are in the same storm but not in the same boat: Impact of COVID-19 on UK households 1444
Demetrio Panarello and Giorgio Tassinari

A network approach to investigate learning experiences and social support in higher education 1450
Ilaria Primerano, Maria Carmela Catone, Giuseppe Giordano, Maria Prosperina Vitale

Physical and cultural activity, internet use and anxiety of Italian university students during the pandemic 1456
Giovanni Busetta, Maria Gabriella Campolo and Demetrio Panarello

The digital divide in Italy before and during the pandemic phase 1462
Laura Zannella

Covid-19 and financial professional advice 1468
Marianna Brunetti and Rocco Ciciretti

Bayesian modelling and inference 4 1472

Bayesian functional mixed effects model for sports data 1473
Patric Dolmeta, Raffaele Argiento and Silvia Montagna

Bayesian Optimization with Machine Learning for Big Data Applications in the Cloud 1479
Bruno Guindani, Danilo Ardagna and Alessandra Guglielmi

Confidence distributions and fusion inference for intractable likelihoods 1485
Elena Bortolato and Laura Ventura

Wasserstein distance and applications to Bayesian nonparametrics 1491
Marta Catalano, Hugo Lavenant, Antonio Lijoi and Igor Prunster

Network Analysis and community detection 1497

Community detection in networks: a heuristic version of Girvan Newman algorithm 1498
Ilaria Bombelli and Lorenzo Di Rocco

Geographically weighted regression for spatial network data: an application to traffic volumes estimation 1504
Andrea Gilardi, Riccardo Borgoni and Jorge Mateu

Asymmetric Spectral Clustering: a comparison between symmetrizations 1510
Cinzia Di Nuzzo and Donatella Vicari

Community detection of seismic point processes 1516
Valeria Policastro, Nicoletta D'Angelo and Giada Adelfio

An Explorative analysis of Different Distance Metrics to Compare Unweighted Undirected Networks 1522
Anna Simonetto, Matteo Ventura and Gianni Gilioli

Gender, attitudes and family ties 1528

Parents of a disabled child in Italy: less healthy but more civically engaged 1529
Nicoletta Balbo and Danilo Bolano

Searching the nexus between women's empowerment and female genital cutting (FGC) 1535
Patrizia Farina, Liva Ortensi, Thomas Pettinato and Enrico Ripamonti

Social stratification, gender, and attitudes towards voluntary childlessness in Europe: A double machine learning approach 1539
Danilo Bolano and Francesco C. Billari

Integrating structuralism and diffusionism to explain the new Italian emigration 1545
Francesca Bitonti

On the effects of rooted family ties in business networks: The South of Italy in the 19th century 1551
Roberto Rondinelli, Giancarlo Ragozini and Maria Carmela Schisani

Methods and Applications in Clustering 1557

A semi-supervised clustering method to extract information from the electronic Word Of Mouth 1558
Giulia Contu, Luca Frigau, Maurizio Romano and Marco Ortu

Spectral approach for clustering three-way data 1564
Cinzia Di Nuzzo and Salvatore Ingrassia

Double clustering with a matrix-variate regression model: finding groups of athletes and disciplines in decathlon's data 1570
Mattia Stival, Mauro Bernardi, Manuela Cattelan and Petros Dellaportas

Classification of the population dynamics 1576
Federico Bacchi and Laura Neri

Locating γ-Ray Sources on the Celestial Sphere via Modal Clustering 1582
Anna Montin, Alessandra R. Brazzale and Giovanna Menardi

Sampling and Official Statistics 1588

Fisher's Noncentral Hypergeometric Distribution for Population Size Estimation 1589
Veronica Ballerini and Brunero Liseo

Small area models for skew and kurtotic distributions 1595
Maria Rosaria Ferrante and Lorenzo Mori

The use of remotely sensed data in sampling designs for forest monitoring 1601
Chiara Bocci, Gherardo Chirici, Giovanni D'Amico, Saverio Francini and Emilia Rocco

Analyzing different causes of one–inflation in capture recapture models for criminal populations 1607
Davide Di Cecco, Andrea Tancredi and Tiziana Tuoto

Administrative database and official statistics: an IT and statistical procedure 1613
Caterina Marini and Vittorio Nicolardi

Spatial modeling and Analyses 1619

Spatial statistics analysis using microdata: an application at agricultural sector 1620
Daniela Fusco, Maria Antonietta Liguori, Valerio Moretti and Francesco Giovanni Truglia

Bayesian spatial modeling of extreme precipitation 1627
Federica Stolf

A proposal to adjust local Moran's I for measuring residential segregation 1632
Antonio De Falco and Antonio Irpino

Accurate directional inference for gaussian graphical models 1637
Claudia Di Caterina, Nancy Reid and Nicola Sartori

Advances in Classification 1643

Measures of interrater agreement based on the standard deviation 1644
Giuseppe Bove

A Comparison of accuracy measures for Classification tasks 1650
Amalia Vanacore and Maria Sole Pellegrino

Iterative Threshold-based Naive Bayes Classifier: an efficient Tb-NB improvement 1656
Maurizio Romano, Gianpaolo Zammarchi and Giulia Contu

Reprogramming FairGANs with Variational Auto-Encoders: A New Transfer Learning Model 1662
Beatrice Nobile, Gabriele Santin, Bruno Lepri and Pierpaolo Brutti

Robust statistics 1669

Combinatorial Analysis of Factorial Designs with Ordered Factors 1670
Roberto Fontana and Fabio Rapallo

Robustifying the Rasch model with the forward search 1676
Anna Comotti and Francesca Greselin

A novel estimation procedure for robust CP model fitting 1682
Valentin Todorov, Violetta Simonacci, Michele Gallo and Nikolay Trendafilov

A robust approach for functional ANOVA with application to additive manufacturing 1688
Fabio Centofanti, Bianca Maria Colosimo, Marco Luigi Grasso, Alessandra Menafoglio, Biagio Palumbo and Simone Vantini

Modeling unconditional M-quantiles in a regression framework 1692
Luca Merlo, Lea Petrella and Nicola Salvati

Model-based clustering 1696

Bayesian mixtures of semi-Markov models 1697
Rosario Barone and Andrea Tancredi

Specification of informative priors for capture-recapture finite mixture models 1703
Pierfrancesco Alaimo Di Loro, Gianmarco Caruso, Marco Mingione, Giovanna Jona Lasinio and Luca Tardella

Clustering multivariate categorical data: a graphical model-based approach 1709
Francesco Rettore, Michele Russo, Luca Zerman and Federico Castelletti

The Gaussian mixture model-based clustering for the comparative analysis of the Healthcare Digitalization Index in the Italian local health authorities 1715
Margaret Antonicelli, Michele Rubino and Filomena Maggino

Student performance evaluation 1721

Rasch model versus Rasch Mixture model: strengthens and limits in identifying factors affecting students' performance in mathematics 1722
Clelia Cascella

Does taking additional Maths classes improve university performance? 1728
Martina Vittorietti, Andrea Priulla and Massimo Attanasio

University dropout and churn in italy: an analysis over time 1734
Barbara Barbieri, Mariano Porcu, Luisa Salaris, Isabella Sulis, Nicola Tedesco and Cristian Usala

The ANOGI for detecting the impact of education and employment on income inequality 1740
Elena Fabrizi, Alessio Guandalini and Alessandra Spagnoli

What causes juvenile crime? a case-control study 1747
Elena Dalla Chiara and Federico Perali

Methods and Applications in Survival analysis 1753

Recursive partitioning for survival data 1754
Ambra Macis

Detecting survival patterns in a digital learning platform 1760
Marta Cannistrà, Mara Soncin and Federico Frattini

An extension of proper Bayesian bootstrap ensemble tree models to survival analysis 1766
Elena Ballante

Modelling time to university dropout by means of time-dependent frailty COX PH models 1771
Mirko Giovio, Paola Mussida and Chiara Masci

Family history in survival and disease development 1777
Maria Veronica Vinattieri and Marco Bonetti

Text mining 1783

Topics & metaverse: an explorative analysis 1784
Emma Zavarrone, Alessia Forciniti, Emanuele Parisi, Maria Gabriella Grassia

Applying Topic Models to bibliographic search: some results in basketball domain 1791
Manlio Migliorati and Eugenio Brentari

Exploiting Text Mining and Network Analysis for future scenarios development: an application on remote working 1797
Yuri Calleo, Simone Di Zio and Vanessa Russo

Emotion recognition in Italian political language to predict positionings and crises government 1803
Alessia Forciniti and Emma Zavarrone

What does your self-description reveal about you? 1809
Riccardo Ricciardi

Variable selection and complete matrix approaches 1815

A Statistical Approach for the Completion of Input-Output Tables 1816
Rodolfo Metulini, Giorgio Gnecco, Francesco Biancalani and Massimo Riccaboni

On multivariate records over sequences of random vectors with Marshall-Olkin dependence of components 1822
A. Khorrami Chokami and Simone A. Padoan

The joint censored gaussian graphical lasso model 1829
Gianluca Sottile, Luigi Augugliaro and Veronica Vinciotti

Variable selection with unbiased estimation: the cdf penalty 1835
Daniele Cuntrera, Vito M.R. Muggeo and Luigi Augugliaro

Automatic variable selection for MIDAS regressions: an application 1841
Consuelo Rubina Nava, Luigi Riso and Maria Grazia Zoia

Distribution Theory and Estimation 1847

A general framework for unit distributions 1848
Francesca Condino, Filippo Domma and Bozidar V. Popovic

Prediction intervals based on multiplicative model combinations 1854
Valentina Mameli and Paolo Vidoni

Some advances on pairwise likelihood estimation in ordinal data latent variable models 1860
Giuseppe Alfonzetti and Ruggero Bellio

Functional Data Analysis 1866

A new functional clustering method: the Functional Clustering and Dimension Reduction model 1867
Adelia Evangelista and Stefano Antonio Gattone

Nonparametric functional prediction bands: theory with an application to bike sharing mobility demand in the city of Milan 1873
Jacopo Diquigiovanni, Matteo Fontana and Simone Vantini

An R package for the statistical process monitoring of functional data 1878
Christian Capezza, Fabio Centofanti, Antonio Lepore, Alessandra Menafoglio, Biagio Palumbo and Simone Vantini

Trend filtering for functional regression 1884
Federico Ferraccioli, Alessandro Casa and Marco Stefanucci

Conformal prediction for spatio-functional regression models 1890
Diana, Romano, Irpino

Tourism and sport studies 1895

Assessing satisfaction of tourists visiting Italian museums: evidence from the eWOM 1896
Daria Mendola and Valentina Oddo

COVID-19 pandemic and tourism demand: a comparison between Spain and Italy 1902
Caterina Sciortino, Ludovica Venturella and Stefano De Cantis

A compositional analysis of tourism in Europe 1908
Francesco Porro

Improving administrative data quality on tourism using Big Data 1914
Antonella Bianchino, Armando d'Aniello and Daniela Fusco

Geographical variations of socio-demographic issues 1920

Elderly HCE and health care need: comparing spatially unexplained levels 1921
Irene Torrini, Laura Rizzi and Luca Grassetti

Measuring sustainable development at the regional level. The case of Italy 1927
Marianna Bartiromo and Enrico Ivaldi

Socio-economic deprivation and COVID-19 infection: a Bayesian spatial modelling approach 1933
Antonino Abbruzzo, Andrea Mattaliano, Alessandro Arrigo, Salvatore Scondotto and Mauro Ferrante

Applications in Economics 1939

The measurement of economic security through relative indicators 1940
Alessandro Gallo, Silvia Pacei and Maria Rosaria Ferrante

A regional analysis of the efficiency by energy's producers in Italy 1946
Gianna Greca, Giuseppe Cinquegrana and Giovanni Fosco

On investigating social and financial aspects of Cardano 1953
Stefano Vacca, Marco Ortu, Gianpaolo Zammarchi and Giuseppe Destefanis

Combined permutation test on the effect of age of micro enterprises on the propensity to Circular Economy 1959
Stefano Bonnini and Michela Borghesi

Comparison of Two Different Approaches to Measure Economic Access to Food and Insecurity: an Application to Mexican data 1965
Stefano Marchetti, Luca Secondi and Adrian Vargas-Lopez

Image analysis and visual methods 1971

Bias correction of the maximum likelihood estimator for Emax model at the interim analysis 1972
Caterina May and Chiara Tommasi

Visual and automated methods in digital microscopy to evaluate fungal colonisation on plant roots 1977
Ivan Sciascia, Andrea Crosino and Andrea Genre

From satellite images to road pavement type: an object-oriented classification approach 1983
Arianna Burzacchi, Matteo Landrò and Simone Vantini

Valid inference for group analysis of functionally aligned fMRI images 1987
Angela Andreella, Riccardo De Santis and Livio Finos

Topological persistence for astronomical image segmentation 1993
Riccardo Ceccaroni, Pierpaolo Brutti, Marco Castellano, Adriano Fontana and Emiliano Merlin

Statistical assessment and empirical estimation 1999

Confidence regions for optimal sensitivity and specificity of a diagnostic test 2000
Gianfranco Adimari, Duc-Khanh To and Monica Chiogna

On the sensitiveness to the memory parameter in the network of tennis 2006
Alberto Arcagni, Vincenzo Candila and Rosanna Grassi

Two-part model with measurement error 2011
Maria Felice Arezzo, Serena Arima, and Giuseppina Guagnano

Statistical assessment of practical significance 2017
Andrea Ongaro, Sonia Migliorati, and Enrico Ripamont

Autoregressive and mixed effects models 2023

Asymptotic Properties of the Nonlinear Least Squares Estimator in HE-HAR Models 2024
Emilija Dzuverovic and Edoardo Otranto

A note on testing for threshold non-linearity in presence of heteroskedasticity in time series 2030
Simone Giannerini and Greta Goracci

The conditional autoregressive Whart-G model 2036
Massimiliano Caporin and Marco Girardi

Semi-parametric generalized linear mixed effects models for binary response for the analysis of heart failure hospitalizations 2042
Alessandra Ragni, Chiara Masci, Francesca Ieva and Anna Maria Paganoni

Issues in Data science 2048

etree: Classification and Regression With Structured and Mixed-Type Data in R 2049
Riccardo Giubilei, Tullia Padellini and Pierpaolo Brutti

Deep Learning framework for ungrouping coarsely aggregated vital rates 2055
Andrea Nigri

Inside the metaverse: analysis of the state of the art and development of a new usage approach based on quality and ethics 2061
Vito Santarcangelo, Emilio Massa, Saverio Gianluca Crisafulli, Antonio Ruoto, Angelo Lamacchia, Alessandro D'Alcantara, Alessandro Verderame and Massimiliano Giacalone

Fisher's Noncentral Hypergeometric Distribution for Population Size Estimation

La distribuzione ipergeometrica noncentrale di Fisher per la stima di numerosit di popolazione

Veronica Ballerini and Brunero Liseo

Abstract Fisher's noncentral hypergeometric (FNCH) distribution naturally suits biased sampling processes. Indeed, this distribution describes a biased urn experiment where balls of different colors are associated with different weights. Despite its potentiality, FNCH distribution has never been applied to official statistics problems, such as the size estimation of heterogeneous populations. Such underuse is mainly due to the computational burden given by its probability mass function, which makes the evaluation of the likelihood function challenging. We present a methodology to estimate the posterior distribution of FNCH parameters, exploiting extra-experimental information and the computational efficiency of MCMC methods. We assess the robustness to weights prior specifications via simulation studies.
Abstract *La distribuzione ipergeometrica non centrale di Fisher (FNCH) si adatta naturalmente a situazioni in cui il processo di campionamento è affetto da distorsione. Tale distribuzione descrive un esperimento di urna distorto, in cui palline di colori diversi sono associate a pesi diversi. Nonostante le sue potenzialit, la FNCH non mai stata applicata a problemi di statistica ufficiale, come la stima della numerosit di popolazioni eterogenee. Tale sottoutilizzo principalmente dovuto all'onere computazionale dato dalla funzione di massa di probabilit, che rende non banale la valutazione della funzione di verosimiglianza. In questo lavoro, presentiamo una metodologia per stimare la distribuzione a posteriori dei parametri della FNCH, sfruttando informazioni extra-sperimentali e l'efficienza computazionale dei metodi MCMC. La robustezza alle diverse specificazioni delle a priori sui pesi valutata tramite studi di simulazione.*

Key words: Official statistics, MCMC, MNAR, Biased sampling

Veronica Ballerini
University of Florence, Piazza di San Marco, 4, 50121 Firenze. e-mail: veronica.ballerini@unifi.it

Brunero Liseo
Sapienza University of Rome, Piazzale Aldo Moro, 5, 00185 Roma. e-mail: brunero.liseo@uniroma1.it

1 Fisher's noncentral hypergeometric distribution

In 2008, Agner Fog clarified the distinction between two distributions, both known in the literature as "the" noncentral hypergeometric distribution (see [7] and [6]). He first solved the nomenclature issue, naming them *Wallenius*' and *Fisher*'s, after the persons who first proposed them (see [5] and [11]).

The main difference between the two distributions resides in the dependence structure of the draws. Assume an urn of size N contains M_1 balls of color 1 and M_2 balls of color 2. The univariate Wallenius' noncentral hypergeometric distribution describes a situation in which the balls are drawn without replacement until n balls are sampled, and the probability to sample X_1 balls of color 1, and X_2 balls of color 2 depends on the colors' relative weights. Such weights can be seen as the probability to retain a ball of that color when drawn (as suggested by [4]). The competing draws make Wallenius' distribution fit perfectly the context of preference or ranking data (see [8]).

Instead, the univariate Fisher's noncentral hypergeometric (FNCH) distribution describes a biased urn experiment when the balls are drawn independently, without replacement, and the sample size n is observed only at the end of the experiment. It is the conditional distribution of two independent Binomial distributions given their sum ([9]):

$$\begin{aligned} X_1 &\sim Binom(M_1, \zeta_1) \\ X_2 &\sim Binom(M_2, \zeta_2) \end{aligned} \tag{1}$$

$$X_1|X_1 + X_2 = n \sim FNCH(M_1, M_2, n, w_1, w_2) \tag{2}$$

where the weights w_c are the odds $\zeta_c/(1-\zeta_c)$, $c = 1, 2$. The weights w_c are defined up to a positive constant k; then, FNCH distribution is identified by the odds ratio $w = w_1/w_2$. Since $M_2 = N - M_1$ and $x_2 = n - x_1$, the formulation (2) is equivalent to:

$$X_1|n \sim FNCH(M_1, N, n, w)\ . \tag{3}$$

Consider ζ_1, ζ_2 in (1) to be the two capture probabilities of the respective two population groups; then, the ratio w may be interpreted as the relative "exposure" of group 1 over group 2 in that capture occasion. As an alternative interpretation, we may consider ζ_1, ζ_2 as the two groups' "non-missing probabilities". In this case, w would be a measure of non-randomness in the missing data mechanism.

Given such considerations, it is clear that FNCH distribution has a great potentiality in the official statistics field. It can manage situations where n units belonging to a heterogeneous population of unknown size N are observed, and such units' capture, or non-missing, probabilities vary according to the population's subgroup they belong to. Such a situation is typical of survey data when the response rates of various groups differ according to their characteristics.

2 Bayesian inference for FNCH

We aim to estimate the size of the heterogeneous population N, or the subgroups' sizes M_1 and M_2. Whether both M_1 and M_2 are unknown, even if w is known and fixed, we can only estimate the relative size of the two groups in the population. Nevertheless, we might have some prior information on one of the groups: such a situation is common when dealing with administrative data. Indeed, consider a sample of resident (group 1) and non-resident (group 2) persons living in a city; given the reliable information contained in the municipal registries, including genuine prior information on M_1 would be legitimate. Moreover, assume data about a sample of self-employed individuals to be available, collecting information on their working situation. Self-employed workers may have (group 1) or have not (group 2) a VAT number. Since people with a VAT number must be listed in social security registers, we still may assume a reasonably concentrated prior on M_1. Finally, imagine we want to estimate the number of unemployed (group 1) and employed (group 2) young Italian graduates of a specific cohort; their respective sizes are unknown. Nevertheless, the National Student Register provides the annual total number of graduates, N; since the associated error is very small, we can elicit a concentrated prior for N.

The subjective elicitation is a debated issue since the attribute "subjective" is often perceived as including personal beliefs in a negative sense. Instead, we consider the elicitation process a rational way to incorporate experts' knowledge and take advantage of their experience; for a deep and detailed discussion about the probabilities' elicitation process, see [3] and [10].

Now, let us assume the following hierarchical model:

$$X_1|X_1 + X_2 = n \sim FNCH(M_1, M_2, n, w) \tag{4}$$

$$M_1, M_2 \overset{\text{ind}}{\sim} \pi(M_c; \boldsymbol{\theta}^{M_c}) \ , c = 1, 2 \tag{5}$$

where $\pi(\cdot)$ denotes a generic distribution depending on some parameters $\boldsymbol{\theta}$.

If also w is unknown, we must elicit a prior on w:

$$w \sim \pi(w; \boldsymbol{\theta}^w) \ . \tag{6}$$

To use a noninformative prior on w is mathematically legitimate. However, we must notice that this parameter plays a crucial role which is often peculiar to the specific data set, and the use of genuine extra-experimental information is strongly suggested here. In the next section, we assess the sensitivity of the posterior to different specifications of w.

We denote with $\pi(M_1, M_2, w)$ the joint prior distribution; it can be factorized into $\pi(M_1)\pi(M_2)\pi(w)$ assuming that the odds ratio for the two groups of being included in the sample is independent on the groups' sizes. We also denote with $L(x_1; M_1, M_2, n, w)$ the likelihood function. Hence, the joint posterior distribution is

$$\pi(M_1, M_2, w|x_1, n) \propto L(x_1; M_1, M_2, n, w)\pi(M_1)\pi(M_2)\pi(w) \ . \tag{7}$$

Table 1: Sensitivity of M_2 to different prior specifications of w. Average variation between the posterior mean of M_2 and the value used to simulate the data, and average standard deviation, estimated on 50 samples for $N = 1000$.

σ_w	Average B_{M2}	Average $sd(\bar{M}_2^*)$
$\sigma_w = 0.2$	0.15	183.2
$\sigma_w = 0.4$	0.17	198.4
$\sigma_w = 1$	0.22	228.3

The posterior in (7) can be easily computed via MCMC methods, e.g., using a Metropolis-within-Gibbs algorithm.

3 Simulation studies

This section shows how the posterior of the subgroups' size changes as the prior specified for the odds becomes wider.

We simulate 50 samples setting $n = 50$ and $w = 2$ for two different population sizes, i.e. $N = 1000$ and $N = 10000$. We assume a very concentrated prior for M_1, namely a Poisson centered on its simulated value, and a discrete Uniform prior distribution for M_2, defined between x_2 and a large upper bound U_{M_2}; we set U_{M_2} equal to 5000 and 15000 for the two population sizes, respectively. Finally, for w we assume a Log-normal prior with parameter $\mu = 2$, and let the standard deviation σ_w vary; see Table 1 and Table 2.

For the two population sizes, Tables 1 and 2 show the average variation between the M_2 posterior and the value we used to simulate data, and the average standard deviation of the M_2 posterior. For each sample, we define our measure of variation as

$$B_{M_2} = \frac{1}{D} \sum_{d=1}^{D} \frac{(M_{2,d} - M_2)}{M_2} \tag{8}$$

where $M_{2,d}$ is d-th draw from the M_2 posterior.

For both population sizes, the mean standard deviation increases as the prior on w becomes wider; moreover, introducing more uncertainty leads to an increasing (upward) mean bias for the posterior mean. This bias is due to the large upper bound we elicit for M_2 prior.

Figures 1 and 2 shows how, however, the posterior is way more concentrated on the true M_2 value than the dashed flat prior (one sample only).

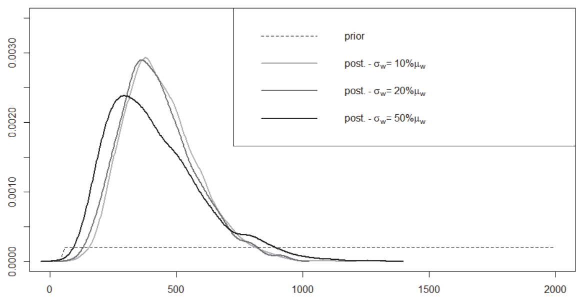

Fig. 1: M_2 prior vs. posteriors with different prior specifications of w. $N = 1000$.

Table 2: Sensitivity of M_2 to different prior specifications of w. Average variation between the posterior mean of M_2 and the value used to simulate the data, and average standard deviation, estimated on 50 samples for $N = 10000$.

σ_w	Average B_{M2}	Average $sd(\bar{M}_2^*)$
$\sigma_w = 0.2$	0.13	1774.2
$\sigma_w = 0.4$	0.17	2051.0
$\sigma_w = 1$	0.29	2818.4

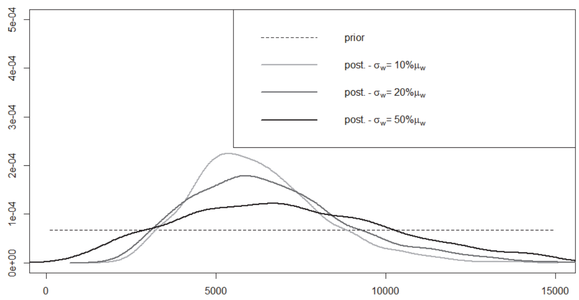

Fig. 2: M_2 prior vs. posteriors with different prior specifications of w. $N = 10000$.

4 Conclusions

Undercoverage is a pervasive issue, mainly addressed in the capture-recapture framework. Although widespread, the single-list case does not boast such vast literature, especially when the target population is elusive — e.g., non-resident inhabitants in a big city, homeless people, irregular migrants. We addressed the problem of size estimation of a heterogeneous population when a single list is available, or we have multiple lists, but we lack unique identifiers. Our model considers the different sources' reliability and the different units' propensity to be captured. Genuine prior information is needed to manage such sources of uncertainty together; a Bayesian model comes naturally in such a situation.

In this short paper, we only discuss the use of our methodology in a simulation study. In [1], an application to the problem of size estimation of unemployed young graduates is presented. The authors exploit the accurate information on the total number of graduates made available by the National Student Register of the Italian Ministry of University and Research. We are currently working to extend the model to the multivariate case. In such a case, the repeated evaluation of the likelihood is prohibitive; thus, a vanilla MCMC algorithm is computationally intractable. As a result, it is needed to use algorithms that avoid the direct evaluation of the FNCH likelihood function. A viable approach is then based on Approximate Bayesian Computation methods, as discussed in [1] and to a more general extent in [2].

References

1. Ballerini, V., Liseo, B.: Fisher's noncentral hypergeometric distribution for the size estimation of unemployed graduates in Italy. Working paper MEMOTEF Sapienza - available upon request (2021)
2. Beaumont, M. A.: Approximate Bayesian Computation. Annual review of statistics and its application **6**, 379-403 (2019)
3. Berger, J. O.: Statistical decision theory and Bayesian analysis. Second Edition. Springer series in statistics, Springer-Verlag (1985)
4. Chesson, J.: A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation. Journal of Applied Probability **13**, 4, 795–797 (1976)
5. Fisher, R. A.: The logic of inductive inference. Journal of the Royal Statistical Society **98**, 1, 39–82 (1935)
6. Fog, A.: Calculation methods for Wallenius' noncentral hypergeometric distribution. Communications in StatisticsSimulation and Computation **37**, 2, 258–273 (2008)
7. Fog, A.: Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. Communications in StatisticsSimulation and Computation **37**, 2, 241–257 (2008)
8. Grazian, C., Leisen, F., Liseo, B.: Modelling preference data with the Wallenius distribution. Journal of the Royal Statistical Society: Series A (Statistics in Society) **182**, 2, 541–558 (2019)
9. Harkness, W. L.: Properties of the extended hypergeometric distribution. The Annals of Mathematical Statistics **36**, 3, 938–945 (1965)
10. O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., Rakow, T.: Uncertain judgements: eliciting experts' probabilities. John Wiley & Sons (2006)
11. Wallenius, K. T.: Biased sampling; the noncentral hypergeometric probability distribution. Stanford University CA Applied Mathematics and Statistics Labs (1963)