

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

BOOK OF SHORT PAPERS

Editors: Rosaria Lombardo, Ida Camminatiello and Violetta Simonacci

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

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



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

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

Program Committee of the conference IES 2022

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

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

Organizing Committee

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



Editors

Rosaria Lombardo - University of Campania “L. Vanvitelli”, Italy

Ida Camminatiello - University of Campania “L. Vanvitelli”, Italy

Violetta Simonacci - University of Naples “L’Orientale”, Italy



PKE - Professional Knowledge Empowerment s.r.l.

Sede legale: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)

Sede operativa: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)

Ufficio Di Rappresentanza: Via Giacomo Peroni, 400 - 00131 Roma (RM)

CF / P.I. 03167830920 — www.pke.it; e-mail info@pke.it — Privacy

February 2022 PKE s.r.l.

ISBN 978-88-94593-35-8 on print

ISBN 978-88-94593-36-5 online

All rights reserved.

This work is protected by copyright law.

All rights, in particular those relating to translation, citation, reproduction in any form, to the use of illustrations, tables and the software material accompanying the radio or television broadcast, the analogue or digital recording, to publication and dissemination through the internet are reserved, even in the case of partial use. The reproduction of this work, even if partial or in digital copy, is admitted only and exclusively within the limits of the law and is subject with the authorization of the publisher. Violation of the rules involves the penalties provided for by the law.

PKE Publisher

Contents

Solicited Session SS1 – <i>Compositional Data Analysis</i>	
Organizer and Chair: Gianna Monti	1
Rieser C. and Filzmoser P. <i>Compositional Data and graph theory</i>	2
Egozcue J. J., Pawlowsky-Glahn V. and Buccianti A. <i>Compositional deviations from linear and non-linear equilibria</i>	8
Monti G. S. and Filzmoser P. <i>The knockoff filter for FDR control in robust ZeroSum regression in microbiome analysis</i>	14
Solicited Session SS2 - <i>Monitoring progress towards SDGs: statistical approaches and methods for measuring poverty, inequalities and food insecurity</i>	
Organizer and Chair: Luca Secondi	18
Marchetti S., Giusti C., Pratesi M. and Biggeri L. <i>Poverty indicators adjusted using local price indexes</i>	19
Tonutti G., Bertarelli G., Giusti C. and Pratesi M. <i>Assessing the targeting of the anti-poverty measure “Reddito di Cittadinanza” using Small Area Estimation methods</i>	25
Vargas-López A. and Secondi L. <i>Household consumption and food insecurity in Mexico: COVID19 and sustainable development</i>	31
Session of free contributes SCL1 – <i>Big Data, Proximity data, Multi way data</i>	
Chair: Donatella Vicari	37
Metulini R. and Carpita M. <i>Forecasting Traffic Flows with Complex Seasonality using Mobile Phone Data</i>	38
Simonacci V., Menini T. and Gallo M. <i>CP decomposition of 4th-order tensors of compositions</i>	44
Bove G. <i>A strategy of analysis of symmetry and skew-symmetry in asymmetric relationships</i>	50
Session of free contributes SCL2 – <i>Industry and Society</i>	

Chair: Germana Scepi	56
Crosato L., Domenech J. and Liberati C. <i>Toward an early detection of SME's default with websites' indicators</i>	57
Angelone R. <i>Italians' Culture and Values after two years of pandemic</i> . .	63
Aria M., Cuccurullo C., D'Aniello L. and Spano M. <i>Thematic evolution of Academic Medical Centers' research: a focus on Italian public owned AOU's in metropolitan areas</i>	67
Session of free contributes SCL3 – Society and Tourism	
Chair: Emma Zavarrone	73
Rondinelli R., Palazzo L. and Ievoli R. <i>Local clustering coefficient to measure intra-regional tourism in Italy</i>	74
Zavarrone E., M. and Forciniti A. <i>Local clustering coefficient to measure intra-regional tourism in Italy</i>	80
Firza N. <i>Sustainable tourism: The case of Albania</i>	86
Romano M., Zammarchi G. and Conversano C. <i>Threshold-based Naïve Bayes Classifier Customer Satisfaction evaluation</i>	90
Session of free contributes SCL4 –Society and Innovation	
Chair: Michelangelo Misuraca	95
Seri E., Alaimo L. S., di Bella E., Cataldo R. and Piscitelli A. <i>Migrant Integration Policy Index (MIPEX): an analysis of countries via Gaussian mixture modelbased clustering</i>	96
Rossi L. and Daddi S. <i>The measure of the BES: a proposal for the aggregation of the indicator education and training</i>	102
Marino M., Mazza R., Misuraca M. and Stavolo A. <i>Monitoring consumer sentiment using control charts</i>	108
Session of solicited contributes SS3 – Statistical methods for the assessment of student careers in higher education	
Organizer and Chair: Maria Prosperina Vitale	114
La Rocca M., Niglio M. and Restaino M. <i>Predicting university students' churn risk</i>	115
Primerano I., Santelli F. and Usala C. <i>Discovering archetypal universities in Italian higher education mobility flows</i>	121
Cascella C. and Ragozzini G. <i>Measuring quality of students' careers in Higher Education: a systematic literature review</i>	127
Porcu M., Sulis I. and Usala C. <i>Estimating the peers effect on students' university choices</i>	134

Session of solicited contributes SS4 – <i>Conformity assessment and quality predictions- itENBIS</i>	
Organizer and Chair: Amalia Vanacore	140
Vanacore A., Pellegrino M. S. and Ciardiello A. <i>Testing the predictive performance of multi-class classifiers</i>	141
Borgoni R., Gilardi A. and Zappa D. <i>Optimal Subgrids from Spatial Monitoring Networks</i>	148
Pennechi F. and Kuselman I. <i>Extension of the JCGM 106:2012 - Conformity assessment of multicomponent items and finite statistical samples</i>	153
 Session of solicited contributes SS5 – <i>Big Data and Large-dimensional Data</i>	
Organizer and Chair: Stefania Mignani	160
Farné M. and Vouldis A. <i>ROBOUT: a conditional outlier detection methodology for large-dimensional data</i>	161
Camillo F. <i>Behaviours, emotions and opinions in modern citizen or customer relationship systems: a correct integration of small and big data for hyper-targeting, personal advertising and look-alike</i>	168
Camminatiello I. and Lucadamo A. <i>A model for assessing sea environmental quality</i>	172
 Session of solicited contributes SS6 – <i>Health Quality</i>	
Organizer and Chair: Paolo Mariani	178
Bartolini B., Bertoldi S., Benedan L., Galeone C., Mariani P., Sofia F. and Zenga M. <i>The uneasiness index in a patient-designed quality of life questionnaire</i>	179
Benedan L., Hachem M. E., Galeone C., Mariani P., Pilo C. and Tadini G. <i>Assessing the Quality of Life of patients with Epidermolysis Bullosa (EB): Development of a patient-centered questionnaire</i>	184
Marletta A. and Morandi M. <i>Survival analysis in a business context: how to control the abandons of my subscribers</i>	190
 Session of free contributes SCL5 – <i>Assessing Performance</i>	
Chair: Cristina Davino	194
di Trapani G. <i>Political performance measuring and tracking through a system based on the Political Performance Indicator (Iep): Naples 2021 case</i>	195
Montanari G. E. and Doretto M. <i>A class of case-mix adjusted probability-based indices for performance evaluation</i>	202

Cavicchia C., Sarnacchiaro P., Vichi M. and Zaccaria G. <i>An ultrametric model to build a Composite Indicators system</i>	208
Session of free contributes SCL6 – Statistical Learning	
Chair: Massimo Aria	212
Migliorati M. and Brentari E. <i>Feature definition for NBA result prediction through Deep Learning</i>	213
Levantesi S., Lizzi M. and Nigri A. <i>An application of contrast trees for mortality models diagnostic and boosting</i>	219
Aria M., Gnasso A. and D’Aniello L. <i>Twenty Years of Random Forest: preliminary results of systematic literature review</i>	225
Park H., Hong J., Shin Y. and Park J-S. <i>A study on the GEV activation function for classification of class imbalance data</i>	231
Session of free contributes SCL7 – Society and Disparity	
Chair: Pasquale Sarnacchiaro	236
Alaimo L. S., D’Urso P. and Nigri A. <i>The gender gap in lifespan disparity as a social indicator of international countries: A fuzzy clustering analysis approach</i>	237
Iannario M. and Tarantola C. <i>Modelling scale effects via a Bayesian approach: an application to decision making in public sector</i>	243
Gangi F. , Daniele L. M. and Coscia M. <i>Board Gender Diversity and Social engagement: evidence from the banking industry</i>	249
Session of free contributes SCL8 – Education	
Chair: Matilde Bini	256
Davino C. and Lamberti G. <i>Assessing heterogeneity in students’ performance. The case of the Massive Open Online Courses</i>	257
Primerano I., Catone M. C., Giordano G. and Vitale M. P. <i>Assessing undergraduate students’ perceptions of distance learning during the COVID-19 pandemic</i>	263
Crisci A., Lucadamo A. and Amenta P. <i>PhD satisfaction analysis in Italian University via Classification tree, Bagging and Random Forest</i>	269
Cervellera S., Cusatelli C. and Giacalone M. <i>Comparative Analysis of Student Learning: Technical, Methodological and Result Assessing of PISA-OECD and INVALSI-Italian Systems</i>	275
Session of solicited contributes SS7 – SEM with PLS: Theory and Applications	
Organizer and Chair: Enrico Ciavolino	281

Wang S., Cheah J. and Roldan J. L. <i>PLS-SEM basics and its potential applications: A quick journey</i>	282
Cefis M. and Carpita M. <i>A PLS-SEM confirmatory composite analysis for football goalkeeper's performance validation</i>	288
Pasca P., Misuraca M., Meloni A. and Ciavolino E. <i>Text-mining and PLS-SEM combination to measure food satisfaction with Google Review: When the gut (re)counts!</i>	294
Session of solicited contributes SS8 – Applications of non standard statistical tools to real-life	
Organizer and Chair: Antonio D'Ambrosio	300
Iorio C. and Pandolfo G. <i>A robust strategy for building a financial portfolio</i>	301
Ruscone M. N. and De Luca G. <i>Conditional copula a financial application</i>	307
Ortu M., Frigau L. and Contu G. <i>Explaining Student Satisfaction Assessments: A Natural Language Processing Approach</i>	313
Session of solicited contributes SS9 – Innovation and Value Co-creation in Society	
Organizers and Chairs: Alessandra De Chiara & Anna D'Auria	319
Mauro S. <i>The smart working towards a Society 5.0</i>	320
Del Vacchio E., Carignani F., Laddaga C. and Bifulco F. <i>Innovative interaction in Society 5.0: insights from the cultural sector</i>	326
D'Auria A. and De Chiara A. <i>Society 5.0: a bibliometric analysis</i>	333
Session of solicited contributes SS10 – Statistical learning for mainstream press, health and fiscal data	
Organizer and Chair: Claudio Conversano	339
Baldassarre A. and Carullo D. <i>The regression trunk model for partitioning Italian municipalities based on their fiscal capacities and its determinants</i>	340
Pandolfo G. and Iorio C. <i>An analysis of Italian healthcare mobility through a depth-based clustering procedure</i>	346
Dossou B. F. P. and Wilhelm A. F. X. <i>Automatic Fake News Detection to Ensure Quality of News Articles</i>	352
Session of solicited contributes SS11 – Multi-way Methods for Evaluation Service	
Organizer and Chair: Michele Gallo	358
Simonacci V., Marino M., Grassia M. G. and Gallo M. <i>Multiple factor analysis with external information on PISA survey data</i>	359

Bocci L. and D. Vicari D. <i>A three-way analysis of well-being in Italy over time</i>	365
Cerqueti R., Mattera R. and Scepi G. <i>Multiway approach for clustering time series with time varying parameters</i>	371
Session of solicited contributes SS12 – Assessment of Management Quality	
Organizer and Chair: Clelia Fiondella	377
Belfiore A., Cuccurullo C. and Aria M. <i>Using partial triadic analysis for depicting the temporal evolution of Italian private healthcare organizations</i>	378
Bollani L., Celegato A., Barbero F. and Fontemaggi F. <i>Management of the human factor into the company. An experience from the aeronautic sector.</i>	385
Session of solicited contributes SS13 – New technologies for students learning assessment and evaluation	
Organizer and Chair: Alfonso Iodice D’Enza	391
Themelis E. and Markos A. <i>A non parametric cognitive diagnostic method in classroom assessment conditions</i>	392
Iannario M., Iodice D’Enza A. and Romano R. <i>Hybrid unfolding models to Likert-scale data to assess distance learning perception in higher education</i>	398
Pacella D., Fabbriatore R., Galluccio C. and Palumbo F. <i>Classification of Statistics learners using multi-dimensional latent class IRT model and archetypal analysis: the ALEAS app</i>	404
Session of solicited contributes SS14 – Labor Market and Enterprises	
Organizer and Chair: Lucio Masserini	408
Maggioni G., Mariani P., Marletta A. and Zenga M. <i>Searching for new trends and dynamics in Labour Market: a statistical approach for the recruiting process</i>	409
Bruttini P., Mariani P., Marletta A., Masserini L. and Zenga M. <i>A new definition of the professional figure Open Manager</i>	413
Session of solicited contributes SS15 – Statistical Approaches to Environmental Sustainability	
Organizer and Chair: Alfonso Piscitelli	417
Aspinall R. <i>Measuring sustainability as an emergent property of whole system dynamics</i>	418

Alaimo L. S. and Finocchiaro G. <i>Tourism sustainability in the Italian regions: a fuzzy approach</i>	424
D'Uggento A. M. <i>How young people perceive environmental issues, react to ecological concerns and commit themselves to sustainable behaviours</i>	430
Session of solicited contributes SS16 – Statistical Methods for Environmental, Natural Resources and Health Assessment	
Organizer and Chair: Alessio Pollice	437
Ferretti A., Ippoliti L. and Valentini P. <i>Spatial-ARFIMA models for the statistical analysis of environmental lattice processes</i>	438
Arima S., Pasculli G. and Polettini S. <i>A Bayesian non parametric approach for bias correction for underreported data</i>	444
Lasinio G. J., Mastrantonio G., Pollice A., Ventura D., Mancini G. and Ardizzone G. <i>Assessment of the impact of anthropic pressures on the Giglio island meadow of Posidonia oceanica</i>	450
Session of solicited contributes SS17 – Functional Data Analysis Methodologies for Quality Assessment	
Organizer and Chair: Elvira Romano	456
Fortuna F., Maturo F. and Di Battista T. <i>Improving the quality of questionnaires via the combined use of functional outlier detection and Item Response Theory</i>	457
Naccarato A., Fortuna F. and Terzi S. <i>Assessing government effectiveness over time: a functional data analysis approach</i>	462
Balzanella A. and Verde R. <i>Mining Distributed Acoustic Sensing data for vehicle traffic monitoring</i>	467
Session of solicited contributes SS18 – Evaluation and assessment of cognitive and learning processes	
Organizer and Chair: Francesco Palumbo	473
Ponticorvo M. and Argiuolo T. <i>Neuropsychological Assessment supported by Technology: the E-BTT case</i>	474
Milano N. and Gigliotta O. <i>Mining Introducing OpenAi-ES in Interactive Data. Clustering with R-EVOK</i>	480
Davino C., Gherghi M., Palumbo F. and Vistocco D. <i>Modeling heterogeneity in student's satisfaction during the Covid-19 pandemia</i> . . .	484
Session of solicited contributes SS19 – Sustainability and Environment	
Organizer and Chair: Ida Camminatiello	490

Lombardo R. and Beh E. J. <i>Partitioning the Cressie-Read divergence statistic for three-way contingency tables: a study on environmental sustainability data</i>	491
Prahadchai T., Hong J., Busababodhin P. and Park J-S.. <i>Analysis of maximum precipitation in Thailand using non-stationary extreme value models</i>	498
Tregua M. and Scaglione M. <i>Assessing citizens' participation to urban transformation: a review of quantitative methods</i>	504
 Session of solicited contributes SS20 – Statistical methods for health and environmental impact assessment	
Organizer and Chair: Fabrizio Maturo	512
Gattone S. A. and Di Battista T. <i>Density estimation via Functional Data Analysis</i>	513
Evangelista A., Acal C., Aguilera A. M., Sarra A., Di Battista T. and Palmeri S. <i>A new multivariate functional ANOVA approach for assessing air quality data amid COVID-19 pandemic</i>	517
Diana A., Romano E. and Irpino A. <i>Conformal Prediction for Geographically Weighted Functional Regression models: an application for environmental impact assessment.</i>	523
 Session of solicited contributes SS21 – Local sustainability assessment: challenges in building quality indicators	
Organizers and Chairs: Francesca Fortuna & Alessia Naccarato	529
Grimaccia E. <i>Urban Sustainability Assessment: A Proposal for an Index Based on SDGs' Indicators</i>	530
Liberati P. and Resce G. <i>Between and Within Country Inequality in Regional Well Being</i>	536
Di Battista T., Nissi E. and Sarra A. <i>Equitable and sustainable well-being over time: a functional approach</i>	542
 Session of free contributes SCL9 – Health and Covid-19	
Chair: Maria Sole Pellegrino	548
Cao N., Calcagni A. and Finos L. <i>Twitter about COVID-19: An application of Structural Topic Models to a sample of Italian tweets</i>	549
Parretti C., Tartaglia R., Sbrana G., Mandó M. and Pacchi S. <i>Assessing the quality of a health service through the risk profile number (RPN)</i>	555
Di Lorenzo G., Franchetti G. and Politano M. <i>The insurance premium structure for a covid-19 insurance policy</i>	562

Session of solicited contributes SS22 – <i>Statistics, culture and tourism</i>	
Organizer and Chair: Marica Manisera	568
Cristiani P. <i>How data can influence the promotion and the consumption of cultural experience</i>	569
Carpita M., Manisera M. and Zuccolotto P. <i>Mobile phone data to monitor the impact of social and cultural events of Brescia</i>	575
Capecchi S., Quaranta G. and Salvia R. <i>Analysing opinions on sustainable tourism in the Vallo di Diano area, Campania, Italy</i>	582
Session of free contributes SCL10– <i>Time Series data, Panel data and Circular Economy</i>	
Chair: Anna Crisci	588
Scaccabarozzi D., Toninelli D., Zurlo D., Bacchini F. and Iannaccone R. <i>Testing the Participation Gap Inclusion within the Wage Phillips Curve</i>	589
Fusco D., Liguori M. A. and Moretti V. <i>Multisource approach for trends evaluation. An application at the agricultural sector</i>	596
Bonnini S. and Borghesi M. <i>A permutation test on the relationship between Circular Economy and firm size</i>	601
Mele S., Izzo F. and Tomnyuk V. <i>Circular economy and business models: a literature review</i>	607
Session of free contributes SCL11 – <i>Modelling Extreme Values, High dimensional, time series data</i>	
Chair: Violetta Simonacci	611
Shin Y., Busababodhin P. and Park J-S. <i>Modeling extreme values using the r-largest four parameter distribution</i>	612
Sabri M., Maturo F., Verde R., Riffi J. and Yahyaouy A. E. <i>Classification of ECG signals based on functional data analysis and machine learning techniques</i>	618
Park J-S., Shin Y., Shin Y. and Hong J. <i>Determining shape parameters in a climate multi-model ensemble</i>	624

A Bayesian non parametric approach for bias correction for underreported data.

Un approccio Bayesiano non parametrico per la correzione dei dati sotto riportati.

Serena Arima, Giuseppe Pasculli and Silvia Poletti

Abstract Data quality is emerging as an essential characteristics of all data driven processes. The implications that data quality issues in computing health or vital statistics may have on government intervention policies and distribution of financial resources highlight the relevance of the problem. In this paper, we deal with the issue of underreporting, paying particular attention to its effects on the estimation of the prevalence of a phenomenon. We propose a non parametric compound Poisson model that allows for the estimation of the reporting probabilities. The proposed model will be applied to a data set concerning early neonatal mortality in Minas Gerais, Brazil. Comparisons of the estimates obtained under several alternative models reveal that the proposed approach is accurate and particularly suitable when there is no prior information about the reporting probability.

Abstract Negli ultimi anni e durante la pandemia, si é confermato il fatto che i dati di qualità sono l'elemento fondamentale per tutti i processi di decisione basati sui dati. Infatti, dati collezionati in modo viziato portano a decisioni distorte con conseguenze su azioni politiche e distribuzioni di fondi. Ciò è particolarmente importante se si tratta di sanità e salute pubblica. In questo lavoro, consideriamo il problema della distorsione dei dati con particolare interesse agli effetti che questo ha sulla stima della prevalenza di un fenomeno. Proponiamo un modello non parametrico di Poisson per la stima delle probabilità di tale distorsione. Il modello è applicato a dati relativi alla mortalitnéonatale in Minas Gerais, Brasile. Le stime verranno confrontate con modelli alternativi e l'applicazione evidenzia che il metodo

Serena Arima

Department of history, society and social sciences, University of Salento, Lecce, Italy, e-mail: serena.arima@unisalento.it;

Giuseppe Pasculli

Department of Computer, Control, and Management Engineering Antonio Ruberti (DIAG), University of Rome "La Sapienza", Rome, Italy, e-mail: giuseppe.pasculli@uniroma1.it ;

Silvia Poletti

Department of Political Sciences, University of Rome "La Sapienza", Rome, Italy, e-mail: silvia.poletti@uniroma1.it

proposto é particolarmente promettente laddove non ci sono informazioni a-priori circa la qualità del dato raccolto.

Key words: Compound Poisson model, hierarchical models, MCMC, Underreporting probabilities,

1 Introduction

In order for scientists to derive reliable statistical analyses, useful for taking appropriate data driven decisions, good quality data are mandatory. As experienced in the last year, inaccurate data collection leads to inappropriate conclusions even when accurate and complex statistical methodologies had been performed. The problem is particularly severe when health or vital statistics are concerned, with important consequences on government intervention policies and distribution of financial resources. In underdeveloped and developing countries, the poor quality of the available data often impairs estimation of economic, health, and social indicators. An example is provided by the case study of this paper, originally presented in [4], where early neonatal mortality risk in Minas Gerais, Brazil, during 1999-2001 is studied. The state of Minas Gerais presents heterogeneous characteristics and a relevant socio-economical inequality. In this context, it is expected that official figures do not reflect the real situation with reports most likely overlooking deaths occurring soon after birth. If such phenomena are not taken into proper account, critical statistics may be underestimated, jeopardizing the concept of effective government intervention policies and the allocation of financial resources.

The problem of underreporting is well known in the literature. The bias caused by a flawed data reporting method is widely discussed in the statistical literature by using hierarchical models that accommodate truncated or censored observations. [4] propose a bias correction method based on a compound Poisson model for count data that includes an area-specific reporting probability, whose uncertainty is accounted for in the model. In the proposal, areas are clustered according to their data quality. Since underreporting probabilities might reflect socio-economic, political and/or demographic characteristics of each region, we focus on modelling the underreporting rates in different areas. Building on the idea in [4], we consider a Bayesian semi-parametric extension of the compound Poisson model: following the approach described in [8], we propose a new method to introduce covariates to define the clustering structure for the underreporting probability ε_i .

2 Modelling underreporting probability

The bias problem induced by defectively reported data is widely discussed in the statistical literature: a great variety of models that accommodate truncated or censored

A Bayesian non parametric approach for bias correction for underreported data.

observations has been developed. Several extensions of the Poisson model have been proposed for counts subject to underreporting ([1], [2], [7]). The compound Poisson model (CPM) is an alternative approach to deal with potentially underreported counts. It allows for the joint modeling of the event occurrence rates and the associated reporting probabilities. [4] discussed issues about the identifiability of the CPM and proposed an alternative Bayesian hierarchical formulation of the CPM that only requires reliable prior information about the reporting process for areas experiencing the best data quality. They rely on the fact that areas within the region under study can be clustered into homogeneous groups reflecting the data quality of each area. The clusters may be defined based on experts' opinion or applying some clustering technique to auxiliary data quality indicators provided by previous studies and surveys.

2.1 A compound Poisson model for underreporting probabilities

Consider a region consisting of m areas and denote by Y_i the observed counts in area i ($i = 1, \dots, m$). Denote by T_i the number of events of interest, and assume

$$Y_i \sim \text{Poisson}(E_i \theta_i),$$

where θ_i is the relative risk and E_i is a known offset representing the expected number of events in the i -th area. T_i is possibly underreported, in the sense that for each record $j = 1, \dots, T_i$ in area i the event is actually reported with probability ε_i , namely $Z_{ij} \sim \text{Bernoulli}(\varepsilon_i)$, $i = 1, \dots, T_i$ independently, and independent on T_i .

The observed counts are defined as $Y_i = \sum_{j=1}^{T_i} Z_{ij}$ and therefore, $Y_i | T_i, \varepsilon_i \sim \text{Bin}(T_i, \varepsilon_i)$. Combining with 2.1 and marginalising over T_i this is the so called compound Poisson model (CPM), under which

$$Y_i | \theta_i \varepsilon_i \sim \text{Poisson}(E_i \theta_i \varepsilon_i).$$

Let us further assume that we can relate the relative risks to a set of covariates $X_1, \dots, X_p$:

$$\log(\theta_i) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}.$$

As described above, the parameter ε_i defines the reporting probability in the i -th area: low values of ε_i indicate areas whose observed counts are underreported. [4] assume that areas can be a-priori clustered according to their data quality. Indeed, they specify

$$\varepsilon_i = 1 - h_i^T \gamma$$

where $(h_{1i}, \dots, h_{Ki})^T$ is the cluster indicator defined according to the following split-coding scheme: if area i belongs to cluster j then $h_{li} = 1$ for all $l \leq j$ and $h_{li} = 0$ otherwise. They also assume that $\gamma \in [0, 1)$ and $\sum_{j=1}^K \gamma_j < 1$ to ensure non-null mean for the associated Poisson distribution.

To ensure identifiability of the model, informative priors about the reporting probabilities and the induced clustering structure among the areas are specified: in particular, [4] fix the number of cluster and model the probability of underreporting according to a-priori knowledge about data quality. Strongly informative priors are defined, especially for the areas that are supposed to be characterized by the best data quality.

However, as revealed by the real data application in [4], parameter estimates do sensibly vary according to the number of clusters and to cluster assignment. Besides this, it might be the case that prior information about data quality is not available. In this work, building on the same compound Poisson model, we consider an alternative approach: we specify a clustering structure for the reporting probability ε_i following a non parametric approach based on a dependent Dirichlet process, that allow the aggregating property of the DP to depend on covariates. Although such a specification is more complex from a theoretical as well as computational point of view, it significantly increases the flexibility of the model since it does not require a-priori knowledge of the number of clusters and it defines the clustering structure in a complete nonparametric, data driven way. Indeed, the clustering is induced by introducing covariates in the stick breaking construction of the Dirichlet process.

3 The proposed model

Let $\varepsilon^n = (\varepsilon_1, \dots, \varepsilon_m)$ and $Z^n = (z_1, \dots, z_m)$ denote, respectively, the entire vector of the reporting probabilities and the covariate Z used as predictor for ε .

A simple nonparametric model can be defined by introducing a DP model on the reporting probabilities:

$$Y_i | \theta_i \varepsilon_i \sim \text{Poisson}(E_i \theta_i \varepsilon_i), \log(\theta_i) = \beta_i + \beta_1 X_{1i} + \dots + \beta_p X_{pi} \quad (1)$$

$$\varepsilon_i | G \sim \text{iid } G \quad (2)$$

$$G | \alpha, G_0 \sim \text{DP}(\alpha, G_0) \quad (3)$$

For any measurable set B , the DP process has the well known stick-breaking representation [9]

$$G(B) = \sum_{j=1}^{\infty} w_j \delta_{\eta_j}(B)$$

where $\delta_{\eta_j}(\cdot)$ is the Dirac measure at η_j and $w_j = V_j \prod_{l < j} [1 - V_l]$ with $V_j | \alpha \stackrel{i.i.d.}{\sim} \text{Beta}(1, \alpha)$ that implies an infinite mixture representation of the distribution of the Y_i .

[3] propose a modification of the well known stick-breaking representation of the DP in which the weights are made dependent on covariates; this is achieved replacing the Beta random variables by normally distributed random variables transformed through the normal cdf. The resulting measure is defined as the probit-stick breaking (PSB) process, see also [9]. As described by [8], [3] allow for dependence on

A Bayesian non parametric approach for bias correction for underreported data.

covariates via the introduction of independent Gaussian processes indexed by the covariates as specified in the following formula:

$$G_z(\cdot) = \sum_{j=1}^{\infty} \left\{ \Phi(\eta_j(z)) \prod_{l < j} [1 - \Phi(\eta_l(z))] \right\} \delta_{\varepsilon_j}(\cdot)$$

where $\eta_j(z) = z' \gamma_j$.

As already discussed in [4], one of the main problems of the CPM is the lack of identifiability. Two main approaches can be used to obtain an identifiable model when sets of covariates, X and H , are used to model the relative risk θ and the reporting probability ε . The first one requires extra information desumed from independent validation datasets and requires non-overlapping and incorrelated X and H . The second one requires informative prior distribution for the overall mean reporting rate which is sufficient to complete the identifiability conditions. In our approach, identifiability is achieved by using two disjoint sets of covariates for estimating the risk and the reporting probability. Moreover, slight informative prior about the intercept parameter β_0 is specified ($\beta_0 \sim N(0, 10)$). Sensitivity analysis with respect to this choice reveal that the model is robust with respect prior specification.

4 Data application and results

We apply our model to data analyzed in [4]: our goal is to accurately estimate the relative risk of early neonatal mortality (ENM) in Minas Gerais State (MG), Brazil, one of the more deprived area of the country. Experts believe that the quality of infant mortality information produced in this country is usually underreported, especially in some areas located in northern and south-Eastern regions of the state. The observed counts of early neonatal deaths have been collected for the $m = 75$ areas from 1999-2001 and 2009-2011. In this work we consider data from 1999-2001. The ENM relative risk assumes a log-linear regression structure which includes local and spatial random effects. Five covariates are introduced in this regression model: the Municipal Human Development Index (MHDI), the proportion of mothers with more than twelve years of formal education (MomEduc), the proportion of children with weight at birth smaller than 2.5 Kg (LowWeight), the proportion of children who were born with some congenital anomaly (Anomaly) and the proportion of mothers who made seven or more prenatal visits during the pregnancy (Prenatal). In [4] the clustering structure is defined according to the quantiles of an adequacy index (AI) introduced by [5] as a measure of the quality of infant mortality data collected in Minas Gerais. In our approach this variable is used in the Dirichlet process for defining both the number of clusters as well as the clustering structure. Informative prior about the intercept parameter is specified according to the obtained estimates with a simple Poisson model.

Table 1 shows the posterior means of the parameters and the corresponding 95% credible intervals. As expected, the two models agree on the fact that only the co-

variate MHDI shows to be significant (likely non-zero effect) to explain the ENM risk for the period 1999–2001. The effect of the covariate MDHI is negative in both periods, indicating that the highest the MHDI, the smallest the ENM risk. According to the approach in [4], the model with $K = 8$ clusters resulted to be the best fitting one while the proposed model selects $K = 5$ clusters, showing a WAIC smaller than the one of the competing model. Also, risk estimates are substantially overlapping confirming that the ENM risks in the poorest areas (North and Northeast) are higher than the ones obtained for more developed regions of Minas Gerais. These results reveal that the proposed approach is accurate and particularly suitable when there are no prior information about data quality.

Table 1 Parameter estimates and 95% credible intervals (in brackets) for the model proposed in [4] and the proposed model.

Parameter	Lopes et al. model $K = 8$ $WAIC = 603.091$	Proposed model $K = 5$ $WAIC = 598.360$
Intercept	1.986 (1.624; 2.348)	0.852 (0.092; 1.612)
MHDI	-3.369 (-3.991; -2.747)	-1.635 (-2.868; -0.402)
MomEduc	-0.033 (-1.263; 1.197)	1.452 (-4.469; 1.317)
LowWeight	-0.095 (-1.279; 1.089)	2.206 (-3.932; 6.836)
Anomaly	2.450 (-11.972; 16.872)	2.207 (-15.644; 20.057)
Prenatal	0.104 (-0.230; 0.528)	0.507 (-0.0311; 1.044)

References

1. Bailey, T. C., Carvalho, M. S., Lapa, T. M., Souza, W. V., and Brewer, M. J. Modeling of under-detection of cases in disease surveillance. *Annals of Epidemiology* **15**(5), 335–343, (2005).
2. Caudill, B. S. and Mixon Jr., F. G. Modeling Household Fertility Decisions: Estimation and Testing of Censored Regression Models for Count Data. *Empirical Economics* **20**(2), 183–196, (1995).
3. Chung, Y. and Dunson, D. B. Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association* **104**, 1646–1660, (2009).
4. de Oliverira, G.L., Argiento, R., Loschi, R.H. Bias correction in clustered underreported data, *Bayesian Analysis TBA*, 1–32, (2020)
5. Franca, E., Abreu, D., Campos, D. and Rausch, M. C. Avaliacao da Qualidade da informacao sobre a mortalidade infantil em Minas Gerais: Utilizacao de uma metodologia simplificada (available in Portuguese). *Revista Medica de Minas Gerais* **16**(1 suppl. 2), 15 –22, ((2006).
6. Hart, J.T. The inverse care law. *Lancet* **7696**, 405–412, (1971).
7. Oliveira, G. L., Loschi, R. H., and Assuncao, R. M. A random-censoring Poisson model for underreported data. *Statistics in Medicine* **36**(30), 4873–4892, (2017).
8. Quintana, F., Mueller, P., Jara, A., MacEachern, S. The dependent Dirichlet process and related models. *arXiv:2007.06129*, (2021)
9. Sethuraman, J.A constructive definition of Dirichlet prior. *Statistica Sinica* **2**, 639–650, (1994).