

Proceedings e report

114

SIS 2017
Statistics and Data Science:
new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference
of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.

(Proceedings e report ; 114)

<http://digital.casalini.it/9788864535210>

ISBN 978-88-6453-521-0 (online)

Peer Review Process

All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Firenze University Press Editorial Board

A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0: <https://creativecommons.org/licenses/by/4.0/legalcode>)

CC 2017 Firenze University Press
Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

SOCIETÀ ITALIANA DI STATISTICA

Sede: Salita de' Crescenzi 26 - 00186 Roma

Tel +39-06-6869845 - Fax +39-06-68806742

email: sis@caspur.it web:<http://www.sis-statistica.it>

La Società Italiana di Statistica (SIS), fondata nel 1939, è una società scientifica eretta ad Ente morale ed inclusa tra gli Enti di particolare rilevanza scientifica. La SIS promuove lo sviluppo delle scienze statistiche e la loro applicazione in campo economico, sociale, sanitario, demografico, produttivo ed in molti altri settori di ricerca.

Organi della società:

Presidente:

- Prof.ssa Monica Pratesi, Università di Pisa

Segretario Generale:

- Prof.ssa Filomena Racioppi, Sapienza Università di Roma

Tesoriere:

- Prof.ssa Maria Felice Arezzo, Sapienza Università di Roma

Consiglieri:

- Prof. Giuseppe Arbia, Università Cattolica del Sacro Cuore

- Prof.ssa Maria Maddalena Barbieri, Università Roma Tre

- Prof.ssa Francesca Bassi, Università di Padova

- Prof. Eugenio Brentari, Università di Brescia

- Dott. Stefano Falorsi, ISTAT

- Prof. Alessio Pollice, Università di Bari

- Prof.ssa Rosanna Verde, Seconda Università di Napoli

- Prof. Daniele Vignoli, Università di Firenze

Collegio dei Revisori dei Conti:

- Prof. Francesco Campobasso, Prof. Michele Gallo, Prof. Francesco Sanna, Prof. Umberto Salinas (supplente)

SIS2017 Committees

Scientific Program Committee:

Rosanna Verde (chair), Università della Campania “Luigi Vanvitelli”
Maria Felice Arezzo, Sapienza Università di Roma
Antonino Mazzeo, Università di Napoli Federico II
Emanuele Baldacci, Eurostat
Pierpaolo Brutti, Sapienza Università di Roma
Marcello Chiodi, Università di Palermo
Corrado Crocetta, Università di Foggia
Giovanni De Luca, Università di Napoli Parthenope
Viviana Egidi, Sapienza Università di Roma
Giulio Ghellini, Università degli Studi di Siena
Ippoliti Luigi, Università di Chieti-Pescara “G. D’Annunzio”
Matteo Mazziotta, ISTAT
Lucia Paci, Università Cattolica del Sacro Cuore
Alessandra Petrucci, Università degli Studi di Firenze
Filomena Racioppi, Sapienza Università di Roma
Laura M. Sangalli, Politecnico di Milano
Bruno Scarpa, Università degli Studi di Padova
Cinzia Viroli, Università di Bologna

Local Organizing Committee:

Alessandra Petrucci (chair), Università degli Studi di Firenze
Gianni Betti, Università degli Studi di Siena
Fabrizio Cipollini, Università degli Studi di Firenze
Emanuela Dreassi, Università degli Studi di Firenze
Caterina Giusti, Università di Pisa
Leonardo Grilli, Università degli Studi di Firenze
Alessandra Mattei, Università degli Studi di Firenze
Elena Pirani, Università degli Studi di Firenze
Emilia Rocco, Università degli Studi di Firenze
Maria Cecilia Verri, Università degli Studi di Firenze

Supported by:

Università degli Studi di Firenze
Università di Pisa
Università degli Studi di Siena
ISTAT
Regione Toscana
Comune di Firenze
BITBANG srl

Index

| | |
|--|-----|
| Preface | XXV |
| Alexander Agapitov, Irina Lackman, Zoya Maksimenko <i>Determination of basis risk multiplier of a borrower default using survival analysis</i> | 1 |
| Tommaso Agasisti, Alex J. Bowers, Mara Soncin <i>School principals' leadership styles and students achievement: empirical results from a three-step Latent Class Analysis</i> | 7 |
| Tommaso Agasisti, Sergio Longobardi, Felice Russo <i>Poverty measures to analyse the educational inequality in the OECD Countries</i> | 17 |
| Mohamed-Salem Ahmed, Laurence Broze, Sophie Dabo-Niang, Zied Gharbi <i>Quasi-Maximum Likelihood Estimators For Functional Spatial Autoregressive Models</i> | 23 |
| Giacomo Aletti, Alessandra Micheletti <i>A clustering algorithm for multivariate big data with correlated components</i> | 31 |
| Emanuele Aliverti <i>A Bayesian semiparametric model for terrorist networks</i> | 37 |

- Giorgio Alleva
Emerging challenges in official statistics: new sources, methods and skills 43
- Rémi André, Xavier Luciani and Eric Moreau
A fast algorithm for the canonical polyadic decomposition of large tensors 45
- Maria Simona Andreano, Roberto Benedetti, Paolo Postiglione, Giovanni Savio
On the use of Google Trend data as covariates in nowcasting: Sampling and modeling issues 53
- Francesco Andreoli, Mauro Mussini
A spatial decomposition of the change in urban poverty concentration 59
- Margaret Antonicelli, Vito Flavio Covella
How green advertising can impact on gender different approach towards sustainability 65
- Rosa Arboretti, Eleonora Carrozzo, Luigi Salmaso
Stratified data: a permutation approach for hypotheses testing 71
- Marika Arena, Anna Calissano, Simone Vantini
Crowd and Minorities: Is it possible to listen to both? Monitoring Rare Sentiment and Opinion Categories about Expo Milano 2015 79
- Maria Felice Arezzo, Giuseppina Guagnano
Using administrative data for statistical modeling: an application to tax evasion 83
- Monica Bailot, Rina Camporese, Silvia Da Valle, Sara Letardi, Susi Osti
Are Numbers too Large for Kids? Possible Answers in Probable Stories 89

| | |
|--|-----|
| Index | IX |
| Simona Balbi, Michelangelo Misuraca, Germana Scepi <i>A polarity-based strategy for ranking social media reviews</i> | 95 |
| A. Balzanella, S.A. Gattone, T. Di Battista, E. Romano, R. Verde <i>Monitoring the spatial correlation among functional data streams through Moran's Index</i> | 103 |
| Oumayma Banouar, Saïd Raghay <i>User query enrichment for personalized access to data through ontologies using matrix completion method</i> | 109 |
| Giulia Barbati, Francesca Ieva, Francesca Gasperoni, Annamaria Iorio, Gianfranco Sinagra, Andrea Di Lenarda <i>The Trieste Observatory of cardiovascular disease: an experience of administrative and clinical data integration at a regional level</i> | 115 |
| Francesco Bartolucci, Stefano Peluso, Antonietta Mira <i>Marginal modeling of multilateral relational events</i> | 123 |
| Francesca Bassi, Leonardo Grilli, Omar Paccagnella, Carla Rampichini, Roberta Varriale <i>New Insights on Students Evaluation of Teaching in Italy</i> | 129 |
| Mauro Bernardi, Marco Bottone, Lea Petrella <i>Bayesian Quantile Regression using the Skew Exponential Power Distribution</i> | 135 |
| Mauro Bernardi <i>Bayesian Factor-Augmented Dynamic Quantile Vector Autoregression</i> | 141 |

- Bruno Bertaccini, Giulia Biagi, Antonio Giusti, Laura Grassini
Does data structure reflect monuments structure? Symbolic data analysis on Florence Brunelleschi Dome
149
- Gaia Bertarelli and Franca Crippa, Fulvia Mecatti
A latent markov model approach for measuring national gender inequality
157
- Agne Bikauskaite, Dario Buono
Eurostat's methodological network: Skills mapping for a collaborative statistical office
161
- Francesco C. Billari, Emilio Zagheni
Big Data and Population Processes: A Revolution?
167
- Monica Billio, Roberto Casarin, Matteo Iacopini
Bayesian Tensor Regression models
179
- Monica Billio, Roberto Casarin, Luca Rossini
Bayesian nonparametric sparse Vector Autoregressive models
187
- Chiara Bocci, Daniele Fadda, Lorenzo Gabrielli, Mirco Nanni, Leonardo Piccini
Using GPS Data to Understand Urban Mobility Patterns: An Application to the Florence Metropolitan Area
193
- Michele Boreale, Fabio Corradi
Relative privacy risks and learning from anonymized data
199
- Giacomo Bormetti, Roberto Casarin, Fulvio Corsi, Giulia Livieri
A stochastic volatility framework with analytical filtering
205

| | |
|---|-----|
| Index | XI |
| Alessandro Brunetti, Stefania Fatello, Federico Polidoro <i>Estimating Italian inflation using scanner data: results and perspectives</i> | 211 |
| Guénael Cabanes, Younès Bennani, Rosanna Verde, Antonio Irpino <i>Clustering of histogram data : a topological learning approach</i> | 219 |
| Renza Campagni, Lorenzo Gabrielli, Fosca Giannotti, Riccardo Guidotti, Filomena Maggino, Dino Pedreschi <i>Measuring Wellbeing by extracting Social Indicators from Big Data</i> | 227 |
| Maria Gabriella Campolo, Antonino Di Pino <i>Assessing Selectivity in the Estimation of the Causal Effects of Retirement on the Labour Division in the Italian Couples</i> | 235 |
| Stefania Capecchi, Rosaria Simone <i>Composite indicators for ordinal data: the impact of uncertainty</i> | 241 |
| Stefania Capecchi, Domenico Piccolo <i>The distribution of Net Promoter Score in socio-economic surveys</i> | 247 |
| Massimiliano Caporin, Francesco Poli <i>News, Volatility and Price Jumps</i> | 253 |
| Carmela Cappelli, Rosaria Simone, Francesca di Iorio <i>Growing happiness: a model-based tree</i> | 261 |
| Paolo Emilio Cardone <i>Inequalities in access to job-related learning among workers in Italy: evidence from Adult Education Survey (AES)</i> | 267 |

- Alessandro Casa, Giovanna Menardi
Signal detection in high energy physics via a semisupervised nonparametric approach
273
- Claudio Ceccarelli, Silvia Montagna, Francesca Petrarca
Employment study methodologies of Italian graduates through the data linkage of administrative archives and sample surveys
279
- Ikram Chairi, Amina El Gonnouni, Sarah Zouinina, Abdelouahid Lyhyaoui
Prediction of Firm's Creditworthiness Risk using Feature Selection and Support Vector Machine
285
- Sana Chakri, Said Raghay, Salah El Hadaj
Contribution of extracting meaningful patterns from semantic trajectories
293
- Chieppa A., Ferrara R., Gallo G., Tomeo V.
Towards The Register-Based Statistical System: A New Valuable Source for Population Studies
301
- Shirley Coleman
Consulting, knowledge transfer and impact case studies of statistics in practice
305
- Michele Costa
The evaluation of the inequality between population subgroups
313
- Michele Costola
Bayesian Non-Negative l_1 -Regularised Regression
319
- Lisa Crosato, Caterina Liberati, Paolo Mariani, Biancamaria Zavarella
Industrial Production Index and the Web: an explorative cointegration analysis
327

| | |
|--|------|
| Index | XIII |
| Francesca Romana Crucinio, Roberto Fontana <i>Comparison of conditional tests on Poisson data</i> | 333 |
| Riccardo D'Alberto, Meri Raggi <i>Non-parametric micro Statistical Matching techniques: some developments</i> | 339 |
| Stefano De Cantis, Mauro Ferrante, Anna Maria Parroco <i>Measuring tourism from demand side</i> | 345 |
| Lucio De Capitani, Daniele De Martini <i>Optimal Ethical Balance for Phase III Trials Planning</i> | 351 |
| Claudia De Vitiis, Alessio Guandalini, Francesca Inglese, Marco D. Terribili <i>Sampling schemes using scanner data for the consumer price index</i> | 357 |
| Ermelinda Della Valle, Elena Scardovi, Andrea Iacobucci, Edoardo Tignone <i>Interactive machine learning prediction for budget allocation in digital marketing scenarios</i> | 365 |
| Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor <i>Nonparametric classification for directional data</i> | 371 |
| Edwin Diday <i>Introduction to Symbolic Data Analysis and application to post clustering for comparing and improving clustering methods by the Symbolic Data Table that they induce</i> | 379 |
| Carlo Drago <i>Identifying Meta Communities on Large Networks</i> | 387 |

- Neska El Haouij, Jean-Michel Poggi, Raja Ghozi, Sylvie Sevestre Ghalila, Mériem Jaidane
Random Forest-Based Approach for Physiological Functional Variable Selection for Drivers Stress Level Classification
393
- Silvia Facchinetti, Silvia A. Osmetti
A risk index to evaluate the criticality of a product defectiveness
399
- Federico Ferraccioli, Livio Finos
Exponential family graphical models and penalizations
405
- Mauro Ferrante, Giovanna Fantaci, Anna Maria Parroco, Anna Maria Milito, Salvatore Scondotto
Key-indicators for maternity hospitals and newborn readmission in Sicily
411
- Ferretti Camilla, Ganugi Piero, Zammori Francesco
Change of Variables theorem to fit Bimodal Distributions
417
- Francesco Finazzi, Lucia Paci
Space-time clustering for identifying population patterns from smartphone data
423
- Annunziata Fiore, Antonella Simone, Antonino Virgillito
IT Solutions for Analyzing Large-Scale Statistical Datasets: Scanner Data for CPI
429
- Michael Fop, Thomas Brendan Murphy, Luca Scrucca
Model-based Clustering with Sparse Covariance Matrices
437
- Maria Franco-Villoria, Marian Scott
Quantile Regression for Functional Data
441

| | |
|--|-----|
| Index | XV |
| Gallo M., Simonacci V., Di Palma M.A. <i>Three-way compositional data: a multi-stage trilinear decomposition algorithm</i> | 445 |
| Francesca Gasperoni, Francesca Ieva, Anna Maria Paganoni, Chris Jackson, Linda Sharples <i>Nonparametric shared frailty model for classification of survival data</i> | 451 |
| Stefano A. Gattone, Angela De Sanctis <i>Clustering landmark-based shapes using Information Geometry tools</i> | 457 |
| Alan E. Gelfand, Shinichiro Shirota <i>Space and circular time log Gaussian Cox processes with application to crime event data</i> | 461 |
| Abdelghani Ghazdali <i>Blind source separation</i> | 469 |
| Massimiliano Giacalone, Antonio Ruoto, Davide Liga, Maria Pilato, Vito Santarangelo <i>An innovative approach for Opinion Mining : the Plutchick analysis</i> | 479 |
| Massimiliano Giacalone, Demetrio Panarello <i>A G.E.D. method for market risk evaluation using a modified Gaussian Copula</i> | 485 |
| Chiara Gigliarano, Francesco Maria Chelli <i>Labour market dynamics and recent economic changes: the case of Italy</i> | 491 |
| Giuseppe Giordano, Giancarlo Ragozini, Maria Prosperina Vitale <i>On the use of DISTATIS to handle multiplex networks</i> | 499 |

- Michela Gnaldi, Silvia Bacci, Samuel Greiff, Thiemo Kunze
Profiles of students on account of complex problem solving (CPS) strategies exploited via log-data
 505
- Michela Gnaldi, Simone Del Sarto
Characterising Italian municipalities according to the annual report of the prevention-of-corruption supervisor: a Latent Class approach
 513
- Silvia Golia
A proposal of a discretization method applicable to Rasch measures
 519
- Anna Gottard
Tree-based Non-linear Graphical Models
 525
- Sara Hbali, Youssef Hbali, Mohamed Sadgal, Abdelaziz El Fazziki
Sentiment Analysis for micro-blogging using LSTM Recurrent Neural Networks
 531
- Stefano Maria Iacus, Giuseppe Porro, Silvia Salini, Elena Siletti
How to Exploit Big Data from Social Networks: a Subjective Well-being Indicator via Twitter
 537
- Francesca Ieva
Network Analysis of Comorbidity Patterns in Heart Failure Patients using Administrative Data
 543
- Antonio Irpino, Francisco de A.T. De Carvalho, Rosanna Verde
Automatic variable and components weighting systems for Fuzzy cmeans of distributional data
 549
- Michael Jauch, Paolo Giordani, David Dunson
A Bayesian oblique factor model with extension to tensor data
 553

| | |
|---|------|
| Index | XVII |
| Johan Koskinen, Chiara Broccatelli, Peng Wang, Garry Robins <i>Statistical analysis for partially observed multilayered networks</i> | 561 |
| Francesco Lagona <i>Copula-based segmentation of environmental time series with linear and circular components</i> | 569 |
| Alessandro Lanteri, Mauro Maggioni <i>A Multiscale Approach to Manifolds Estimation</i> | 575 |
| Tiziana Laureti, Carlo Ferrante, Barbara Dramis <i>Using scanner and CPI data to estimate Italian sub-national PPPs</i> | 581 |
| Antonio Lepore <i>Graphical approximation of Best Linear Unbiased Estimators for Extreme Value Distribution Parameters</i> | 589 |
| Antonio Lepore, Biagio Palumbo, Christian Capezza <i>Monitoring ship performance via multi-way partial least-squares analysis of functional data</i> | 595 |
| Caterina Liberati, Lisa Crosato, Paolo Mariani, Biancamaria Zavanella <i>Dynamic profiling of banking customers: a pseudo-panel study</i> | 601 |
| Giovanni L. Lo Magno, Mauro Ferrante, Stefano De Cantis <i>A comparison between seasonality indices deployed in evaluating unimodal and bimodal patterns</i> | 607 |
| Rosaria Lombardo, Eric J Beh <i>Three-way Correspondence Analysis for Ordinal-Nominal Variables</i> | 613 |

- Monia Lupparelli, Alessandra Mattei
Log-mean linear models for causal inference
621
- Badiaa Lyoussi, Zineb Selihi, Mohamed Berraho, Karima El Rhazi, Youness El Achhab, Adiba El Marrakchi, Chakib Nejjari
Research on the Risk Factors accountable for the occurrence of degenerative complications of type 2 diabetes in Morocco: a prospective study
627
- Valentina Mameli, Debora Slanzi, Irene Poli
Bootstrap group penalty for high-dimensional regression models
633
- Stefano Marchetti, Monica Pratesi, Caterina Giusti
Improving small area estimates of households' share of food consumption expenditure in Italy by means of Twitter data
639
- Paolo Mariani, Andrea Marletta, Mariangela Zenga
Gross Annual Salary of a new graduate: is it a question of profile?
647
- Maria Francesca Marino, Marco Alfò
Dynamic random coefficient based drop-out models for longitudinal responses
653
- Antonello Maruotti, Jan Bulla
Hidden Markov models: dimensionality reduction, atypical observations and algorithms
659
- Chiara Masci, Geraint Johnes, Tommaso Agasisti
A flexible analysis of PISA 2015 data across countries, by means of multilevel trees and boosting
667

| | |
|---|-----|
| Index | XIX |
| Lucio Masserini, Matilde Bini <i>Impact of the 2008 and 2012 financial crises on the unemployment rate in Italy: an interrupted time series approach</i> | 673 |
| Angelo Mazza, Antonio Punzo, Salvatore Ingrassia <i>An R Package for Cluster-Weighted Models</i> | 681 |
| Antonino Mazzeo, Flora Amato <i>Methods and applications for the treatment of Big Data in strategic fields</i> | 687 |
| Letizia Mencarini, Viviana Patti, Mirko Lai, Emilio Sulis <i>Happy parents' tweets</i> | 693 |
| Rodolfo Metulini, Marica Manisera, Paola Zuccolotto <i>Space-Time Analysis of Movements in Basketball using Sensor Data</i> | 701 |
| Giorgio E. Montanari, Marco Doretto, Francesco Bartolucci <i>An ordinal Latent Markov model for the evaluation of health care services</i> | 707 |
| Isabella Morlini, Maristella Scorza <i>New fuzzy composite indicators for dyslexia</i> | 713 |
| Fionn Murtagh <i>Big Textual Data: Lessons and Challenges for Statistics</i> | 719 |
| Gaetano Musella, Gennaro Punzo <i>Workers' skills and wage inequality: A time-space comparison across European Mediterranean countries</i> | 731 |

- Marta Nai Ruscone
Exploratory factor analysis of ordinal variables: a copula approach 737
- Fausta Ongaro, Silvana Salvini
IPUMS Data for describing family and household structures in the world 743
- Tullia Padellini, Pierpaolo Brutti
Topological Summaries for Time-Varying Data 747
- Sally Paganin
Modeling of Complex Network Data for Targeted Marketing 753
- Francesco Palumbo, Giancarlo Ragozini
Statistical categorization through archetypal analysis 759
- Michela Eugenia Pasetto, Umberto Noè, Alessandra Luati, Dirk Husmeier
Inference with the Unscented Kalman Filter and optimization of sigma points 767
- Xanthi Pedeli, Cristiano Varin
Pairwise Likelihood Inference for Parameter-Driven Models 773
- Felicia Pelagalli, Francesca Greco, Enrico De Santis
Social emotional data analysis. The map of Europe 779
- Alessia Pini, Lorenzo Spreafico, Simone Vantini, Alessandro Vietti
Differential Interval-Wise Testing for the Inferential Analysis of Tongue Profiles 785
- Alessia Pini, Aymeric Stamm, Simone Vantini
Hotelling meets Hilbert: inference on the mean in functional Hilbert spaces 791

| | |
|--|-----|
| Index | XXI |
| Silvia Poletti, Serena Arima <i>Accounting for measurement error in small area models: a study on generosity</i> | 795 |
| Gennaro Punzo, Mariateresa Ciommi <i>Structural changes in the employment composition and wage inequality: A comparison across European countries</i> | 801 |
| Walter J. Radermacher <i>Official Statistics 4.0 – learning from history for the challenges of the future</i> | 809 |
| Fabio Rapallo <i>Comparison of contingency tables under quasi-symmetry</i> | 821 |
| Valentina Raponi, Cesare Robotti, Paolo Zaffaroni <i>Testing Beta-Pricing Models Using Large Cross-Sections</i> | 827 |
| Marco Seabra dos Reis, Biagio Palumbo, Antonio Lepore, Ricardo Rendall, Christian Capezza <i>On the use of predictive methods for ship fuel consumption analysis from massive on-board operational data</i> | 833 |
| Alessandra Righi, Mauro Mario Gentile <i>Twitter as a Statistical Data Source: an Attempt of Profiling Italian Users Background Characteristics</i> | 841 |
| Paolo Righi, Giulio Barcaroli, Natalia Golini <i>Quality issues when using Big Data in Official Statistics</i> | 847 |
| Emilia Rocco <i>Indicators for the representativeness of survey response as well as convenience samples</i> | 855 |

- Emilia Rocco, Bruno Bertaccini, Giulia Biagi, Andrea Giommi
A sampling design for the evaluation of earthquakes vulnerability of the residential buildings in Florence
861
- Elvira Romano, Jorge Mateu
A local regression technique for spatially dependent functional data: an heteroskedastic GWR model
867
- Eduardo Rossi, Paolo Santucci de Magistris
Models for jumps in trading volume
873
- Renata Rotondi, Elisa Varini
On a failure process driven by a self-correcting model in seismic hazard assessment
879
- M. Ruggieri, F. Di Salvo and A. Plaia
Functional principal component analysis of quantile curves
887
- Massimiliano Russo
Detecting group differences in multivariate categorical data
893
- Michele Scagliarini
A Sequential Test for the C_{pk} Index
899
- Steven L. Scott
Industrial Applications of Bayesian Structural Time Series
905
- Catia Scricciolo
Asymptotically Efficient Estimation in Measurement Error Models
913

| | |
|---|-------|
| Index | XXIII |
| Angela Serra, Pietro Coretto, Roberto Tagliaferri <i>On the noisy high-dimensional gene expression data analysis</i> | 919 |
| Mirko Signorelli <i>Variable selection for (realistic) stochastic blockmodels</i> | 927 |
| Marianna Siino, Francisco J. Rodriguez-Cortés, Jorge Mateu, Giada Adelfio <i>Detection of spatio-temporal local structure on seismic data</i> | 935 |
| A. Sottosanti, D. Bastieri, A. R. Brazzale <i>Bayesian Mixture Models for the Detection of High-Energy Astronomical Sources</i> | 943 |
| Federico Mattia Stefanini <i>Causal analysis of Cell Transformation Assays</i> | 949 |
| Paola Stolfi, Mauro Bernardi, Lea Petrella <i>Estimation and Inference of SkewStable distributions using the Multivariate Method of Simulated Quantiles</i> | 955 |
| Paola Stolfi, Mauro Bernardi, Lea Petrella <i>Sparse Indirect Inference</i> | 961 |
| Peter Struijs, Anke Consten, Piet Daas, Marc Debusschere, Maiki Ilves, Boro Nikic, Anna Nowicka, David Salgado, Monica Scannapieco, Nigel Swier <i>The ESSnet Big Data: Experimental Results</i> | 969 |
| Jérémie Sublime <i>Smart view selection in multi-view clustering</i> | 977 |

- Emilio Sulis
Social Sensing and Official Statistics: call data records and social media sentiment analysis
985
- Matilde Trevisani, Arjuna Tuzzi
Knowledge mapping by a functional data analysis of scientific articles databases
993
- Amalia Vanacore, Maria Sole Pellegrino
Characterizing the extent of rater agreement via a non-parametric benchmarking procedure
999
- Maarten Vanhoof, Stephanie Combes, Marie-Pierre de Bellefon
Mining Mobile Phone Data to Detect Urban Areas
1005
- Viktoriya Voytsekhovska, Olivier Butzbach
Statistical methods in assessing the equality of income distribution, case study of Poland
1013
- Ernst C. Wit
Network inference in Genomics
1019
- Dilek Yildiz, Jo Munson, Agnese Vitali, Ramine Tinati, Jennifer Holland
Using Twitter data for Population Estimates
1025
- Marco Seabra dos Rei
Structured Approaches for High-Dimensional Predictive Modeling
1033

Preface

The 2017 SIS Conference aims to highlight the crucial role of the Statistics in Data Science. In this new domain of “meaning” extracted from the data, the increasing amount of produced and available data in databases, nowadays, has brought new challenges. That involves different fields of statistics, machine learning, information and computer science, optimization, pattern recognition. These afford together a considerable contribute in the analysis of “Big data”, open data, relational and complex data, structured and no-structured. The interest is to collect the contributes which provide from the different domains of Statistics, in the high dimensional data quality validation, sampling extraction, dimensional reduction, pattern selection, data modelling, testing hypotheses and confirming conclusions drawn from the data. In the mention that statistics is the “grammar of data science”, statistics has become a basic skill in data science: it gives right meaning to the data. Still, it isn’t replaced by newer techniques from machine learning and other disciplines but it complements them. The Conference is also addressed to the new challenges of the new generations: the native digital generations, who are called to develop professional skills as “data analyst”, one of the more request professionalism of the 21st Century, crossing the rigid disciplinary domains of competence. In this perspective, all the traditional statistical topics are admitted with an extension to the related machine learning and computer science ones. The present volume includes the short papers of the contributions that will be presented in the 4 invited speaker sessions; in the 19 specialized sessions; in the 11 solicited sessions; in the 6 foreign societies sessions and in the 17 contributed sessions as well as, in the panel session.

Rosanna Verde
President of the Scientific Programme Committee

Alessandra Petrucci
President of the Local Organizing Committee

Accounting for measurement error in small area models: a study on generosity.

Modelli per piccola area con errore di misurazione: uno studio sulla generosità

Silvia Poletti and Serena Arima

Abstract In this paper we focus on a recently documented effect of economic inequality, namely that higher income individuals tend to be less generous than poorer individuals, *but only in contexts where macro-level economic inequality is high*, or is perceived as high. We consider data from the Measuring Morality study, a nationally representative survey of United States residents, that contains a validated behavioural measure of generosity (the dictator game) along with the household income of respondents. We fit a small area model to this data with the aim of investigating the role of economic inequality on generosity in the US. We observe that model covariates (reported income and Gini index) are subject to measurement error and investigate the effect of introducing the measurement error in this model.

Abstract *Il lavoro considera il ruolo della disuguaglianza economica sulla generosità, a partire da uno studio recente secondo cui gli individui con redditi più elevati tendono ad essere meno generosi degli individui meno abbienti, ma solo in contesti di grande disuguaglianza economica. I dati analizzati provengono dal Measuring Morality study, un'indagine effettuata negli USA in cui viene rilevato il reddito e una misura validata di generosità (dictator game). Per ogni area di residenza è stato anche ricavato l'indice di Gini, come misura di disuguaglianza economica. In questo lavoro si stima la generosità mediante un modello per piccole aree con reddito e disuguaglianza come variabili ausiliarie. Il modello viene esteso al fine di considerare l'errore di misurazione nelle variabili ausiliarie, sia continue che discrete.*

Key words: small area estimation, measurement error, misclassification, Bayesian inference.

Silvia Poletti

Dip. di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza Università di Roma, via del Castro Laurenziano, 9, e-mail: silvia.poletti@uniroma1.it

Serena Arima

Dip. di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza Università di Roma, via del Castro Laurenziano, 9, e-mail: serena.arima@uniroma1.it

1 Introduction

There is an increasing interest in understanding the implications of income for behaviour, in particular generosity toward others. Well grounded literature on this topic has portrayed a picture of higher-income individuals as consistently more selfish than poorer individuals [13]. A different perspective is reported in a recent paper [6], where the relationship between economic inequality, income, and generosity is tested. Analysing data from the Measuring Morality study (a nationally representative survey of United States residents), as well as a follow-up experiment, the authors identify a previously undocumented effect of economic inequality, namely that higher income individuals in the US tend to be less generous than poorer individuals, *but only in contexts where macro-level economic inequality is high*, or is perceived as high. The Authors comment that the results obtained challenge the prevailing view in the literature that higher income individuals are necessarily less generous and conclude that “inequitable resource distributions undermine collective welfare” and that redistributive policies may “attenuate, or even reverse, the negative relationship between income and generosity, in turn increasing the generosity of those individuals who have the most to give”.

The Measuring Morality study data contain a validated behavioural measure of generosity (the dictator game) along with the household income of respondents; moreover, Gini indices were available from the American Community Survey. The authors fit a mixed effects model to these data, where significant, negative, interaction between income and inequality is found. Using a Bayesian approach, we consider the same model, in a small area context and speculate on the fact that both income and the Gini index are subject to measurement error for different reasons: indeed income is self reported and the Gini index is estimated from another survey. As stressed in the literature, ignoring the measurement error in the covariates may lead to inconsistent estimates and can severely invalidate inferences.

The paper is organized as follows: in Section 2 we introduce the problem of measurement error in small area estimation and propose a small area model accounting for measurement error in covariates and present. In Section 3 we present and discuss the results obtained when the model is applied to the generosity data.

2 A measurement error small area model for generosity data

In this paper, we focus on unit level small area models, within a Bayesian framework. Unit level small area models relate the unit values of the study variable to unit-specific auxiliary variables with known area means. See [11] for an up-to-date review.

Suppose there are m areas and let N_i be the known population size of area i . We denote by Y_{ij} the response of the j -th unit in the i -th area ($i = 1, \dots, m$; $j = 1, \dots, N_i$). A random sample of size n_i is drawn from the i -th area. The goal is to predict the small area means $\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}$, $i = 1, \dots, m$, based on the available

data. To develop reliable estimates, auxiliary information is introduced as covariates and usually a mixed effects model is specified as

$$Y_{ij} = \alpha + \beta w_{ij} + u_i + \varepsilon_{ij} \quad i = 1, \dots, m; \quad j = 1, \dots, N_i \quad (1)$$

with ε_{ij} and u_i independent, $\varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$ and $u_i \stackrel{iid}{\sim} N(0, \sigma_u^2)$. [8] and [9] were the first to consider the problem of measurement error in small area models for unit-level data. They assume that the true, area-level, covariate, w_i , is measured with error as

$$S_{ij} = w_i + \eta_{ij}, \quad \eta_{ij} \stackrel{iid}{\sim} N(0, \sigma_\eta^2) \quad i = 1, \dots, m; \quad j = 1, \dots, n_i \quad (2)$$

where ε_{ij} , u_i and η_{ij} are taken mutually independent. [8] also assumed that $w_i \stackrel{iid}{\sim} N(\mu_w, \sigma_w^2)$, defining the structural measurement error model. They considered both an empirical Bayes and a hierarchical Bayes approach to derive predictors of small area means θ_i . [12] extended the approach in [8] including sample information on the covariate values. [8] also proposed a fully Bayesian approach, by specifying a hierarchical model, with vague prior distributions for all the model parameters, whose posterior distributions are estimated via Gibbs sampling. [1, 3] extended the above approach, proposing to use the Jeffreys' prior on the model parameters. The aforementioned literature considers the case in which the measurement error only affects continuous variables, according to the measurement error model of equation (1). For discrete covariates, measurement error means misclassification. To allow for auxiliary discrete covariates measured with error, [4] propose to model the misclassification mechanism through an unknown transition matrix P and estimate all the unknown parameters in a fully Bayesian framework. Following [4], for each unit in each area, we consider the following covariates: t_{ij} – the vector of p continuous or discrete covariates measured without error, w_i and x_{ij} – respectively, a vector of q continuous covariates and h discrete variables (with a total of K categories), both measured with error. Denote by s_{ij} and z_{ij} the observed values of the latent w_i and x_{ij} , respectively. Without loss of generality, in what follows we assume $h = 1$.

Following the notation in [8], the proposed measurement error model can be written in the usual multi-stage way: for $j = 1, \dots, n_i$, $i = 1, \dots, m$ and for $k, k' = 1, \dots, K$

- Stage 1. $y_{ij} = \theta_{ij} + e_{ij}$ $e_{ij} \stackrel{iid}{\sim} N(0, \sigma_e^2)$
- Stage 2. $\theta_{ij} = t'_{ij}\delta + w'_i\gamma + \sum_{k=1}^K I(x_{ij} = k)\beta_k + u_i$ $u_i \stackrel{iid}{\sim} N(0, \sigma_u^2)$
- Stage 3. $S_{ij}|w_i \stackrel{iid}{\sim} N(w_i, \Sigma_s = \text{diag}(\sigma_{s_1}^2, \dots, \sigma_{s_q}^2))$ $W_i \stackrel{iid}{\sim} N(0, \Sigma_w = \text{diag}(\sigma_{w_1}^2, \dots, \sigma_{w_q}^2))$
 $\Pr(Z_{ij} = k|X_{ij} = k') = p_{k'k}$, $p_{k'k} \sim \text{Dir}(\alpha_{k',1}, \dots, \alpha_{k',K})$ $\Pr(X_{ij} = k') = \frac{1}{K}$
- Stage 4. β , δ , γ , σ_e^2 , σ_u^2 , $\sigma_{s_1}^2, \dots, \sigma_{s_p}^2$ are, loosely speaking, a-priori mutually independent.

Stage 3 defines the measurement error model for both continuous and discrete covariates. For the discrete covariates, the misclassification mechanism is specified according to the $K \times K$ matrix P , whose (k', k) element, $p_{k'k}$, denotes the probability

that the observable variable Z_{ij} takes the k -th category when the true unobservable variable X_{ij} takes the k' -th category. We also assume that the misclassification probabilities are the same across subjects and that all the categories have the same prior probability $\frac{1}{K}$ to occur. Over each row of P , we place a Dirichlet $Dir(\alpha_{k',1}, \dots, \alpha_{k',K})$ prior distribution, with known $\alpha_{k',1}, \dots, \alpha_{k',K}$. In Stage 4 we assume Normal priors for β , δ , and γ and inverse gamma distributions for σ_e^2 and σ_u^2 and σ_s^2 . Hyperparameters have been chosen to have flat priors. Finally, we fix Σ_w and $(\alpha_{k',1}, \dots, \alpha_{k',K})$. According to the above assumptions, we can estimate the transition matrix P and the measurement error variance σ_s^2 jointly with all the other model parameters. As the posterior distribution cannot be derived analytically in closed form, we obtain samples from the posterior distribution using Gibbs sampling.

3 Results and conclusions

We fit a unit level small area model with measurement error in covariates, which also allows us to evaluate the relationship between economic inequality, income and generosity. We use data from the Measuring Morality study, a nationally representative survey of United States residents consisting of a sample of 1498 respondents in the US. For each respondent, income and some personal and demographic variables (such as age, gender, education, ...) have been collected. Respondents completed a validated behavioural measure of generosity: the dictator game. Respondents learned that they had been randomly assigned the role of *decider* and had received 10 tickets, each worth one entry in a raffle to win a monetary prize of either 10 or 500. They could transfer any number of tickets to the next participant, a *receiver* who did not have any tickets. By giving tickets, respondents could benefit another person at a cost to themselves in a zero-sum opportunity to win money. This measure of generosity was administered to individuals with different incomes residing in areas (US states plus the District of Columbia) that vary in levels of inequality, measured according to the Gini's coefficient. The number of respondents in each area ($m = 9$ divisions) ranges from 72 to 286. In the proposed model we take generosity as the response variable and income, standardized Gini coefficients and their interaction as auxiliary variables. According to the survey design, household income was collected as a 19-classes variable; for ease of interpretation in the application we recoded it into five classes (C_1 : less than 12500; C_2 : [12500, 30000), C_3 : (30000, 60000], C_4 : (60000, 125000], C_5 : over 125000). Since income is self reported and the Gini index is estimated using data from the 2012 American Community survey, we can suspect that both auxiliary variables are subject to measurement error. In order to evaluate the impact of accounting for this source of error, we fit both the standard model that ignores the measurement error and the model proposed in Section 2. Figure 1 shows the posterior distribution of the model parameters. The left panel reports the posterior distribution of the regression parameters under the proposed measurement error model: income is the only factor that significantly impacts on the response variable, since for all the other pa-

rameters the 95% credible intervals contain the zero value ($CI_{Gini} : [-0.207, 0.349]$, $CI_{C1*Gini} : [-0.632, 0.241]$, $CI_{C2*Gini} : [-0.542, 0.217]$, $CI_{C3*Gini} : [-0.533, 0.189]$, $CI_{C4*Gini} : [-0.827, -0.028]$). With respect to the income, it is apparent that generosity increases with income, with the exception of the last class, in which the effect on generosity is comparable to that of the second one. This actually means that the richest are less generous with respect to the others, which is line with findings in the mainstream literature on the subject. On the other hand, when one ignores the measurement error, all the covariates and their interactions seem to be significant (Figure 1, right panel). In particular, income exhibits a positive effect on generosity, with no distinctions between income classes, which contradicts the economic theories; moreover, an unexpectedly positive effect of inequality is found. With respect to the measurement error for income, the posterior distribution of $P_{1,1}$ is concentrated around 0.5 and almost uniformly distributed over the other categories. This is an empirical evidence that income is often underreported by the respondents. The distributions of the other diagonal elements of P are concentrated around 0.9 and credible intervals do not contain 1. We conclude that measurement error has a significant impact on income. The small area estimates produced under the model with and without measurement error are reported in Table 1. As can be seen, allowing for measurement error in both continuous and categorical covariates also impacts on estimation of the small area means in both point estimates (in particular for the first division, which is one of the smallest ones) and measures of uncertainty. Also, although the posterior means are not very different for the large areas, the ranking of the divisions varies. As can be seen, allowing for measurement error in both continuous and categorical covariates also impacts on estimation of the small area means. Although the posterior means are not very different, the ranking of the divisions varies. In conclusion, our application reveals that ignoring the measurement error in covariates may drive inferences and yield misleading conclusions.

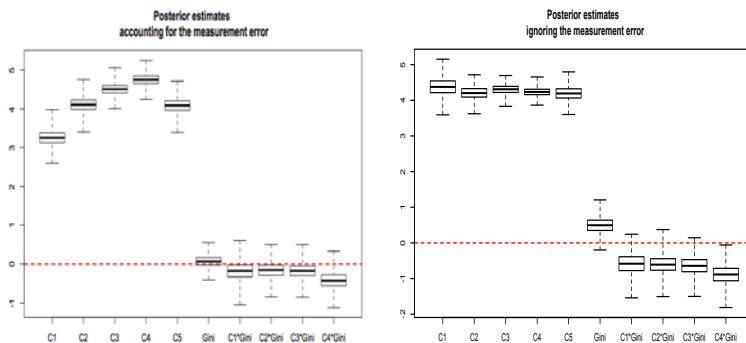
Table 1 Small area estimates: posterior means of the small area means obtained with the model that does not account for the measurement error (first row) and the model that accounts for it (second row). Standard deviations in brackets.

| Division | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| θ_{NoErr} | 4.17 (0.27) | 4.11 (0.33) | 4.25 (0.18) | 4.44 (0.20) | 4.19 (0.24) | 4.28 (0.10) | 4.25 (0.14) | 4.37 (0.16) | 4.22 (0.23) |
| θ_{Err} | 4.27 (0.36) | 4.09 (0.41) | 4.26 (0.38) | 4.43 (0.37) | 4.17 (0.40) | 4.30 (0.33) | 4.25 (0.34) | 4.38 (0.32) | 4.23 (0.40) |

References

1. Arima, S., Datta, G.S., Liseo, B.: Objective Bayesian analysis of a measurement error small area model. *Bayesian Analysis*, **72** (2),363–384 , (2012)

Fig. 1 Posterior distribution of the model parameters. Left panel: posterior distributions obtained from the proposed model. Right panel: posterior distributions from the model that ignores the measurement error.



2. Arima, S., Datta, G.S., Liseo, B.: Bayesian Estimators for Small Area Models when Auxiliary Information is Measured with Error. *Scandinavian Journal of Statistics*, **42** (2),518–529, (2014)
3. Arima, S., Datta, G.S., Liseo, B.: Models in Small Area Estimation when Covariates are Measured with Error, in *Analysis of Poverty Data by Small Area Estimation*, 151–170, (2015)
4. Arima, S., Poletini, S.: A unit-level small area model with misclassified covariates, arXiv:1611.02845 [stat.ME],(2016)
5. Carroll, R.J., Ruppert, D., Stefanski, L., Crainiceanu, C.: *Measurement error in nonlinear models: a modern perspective*. 2nd edn. Chapman & Hall, CRC, (2006)
6. Côté, S., House, J., Willer, R.: High economic inequality leads higher-income individuals to be less generous . *Ann. PNAS*, **112**, 52, 15838–15843 (2015)
7. Engel, C.: Dictator Games: A Meta Study. *Experimental Economics* **14**(4), 583?-610, (2011)
8. Ghosh, M., Sinha, K. and Kim, D.: Empirical and Hierarchical Bayesian estimation in finite population sampling under structural measurement error model. *Scandinavian Journal of Statistics*, **33**(3), (2006)
9. Ghosh, M., Sinha, K.: Empirical Bayes estimation in finite population sampling under functional measurement error models. *Journal of Statistical Planning Inference*,**137**, 2759–2773,(2007)
10. Poletini, S., Arima, S.: Small area estimation with covariates perturbed for disclosure limitation. *Statistica*, **25** (1), 57–72, (2015)
11. Rao, J.N.K. and Molina, I.: *Small Area Estimation*, 2nd Edition, Wiley, Hoboken, New Jersey, (2015).
12. Torabi, M., Datta, G.S. and Rao, J.N.K. Empirical Bayes estimation of small area means under nested error linear regression model with measurement error in the covariates, *Scandinavian Journal of Statistics*, **36**, 355–368, (2009).
13. Trautmann, S.T., van de Kuilen, G. and Zeckhauser, R.J.: Social class and (un)ethical behavior: A framework, with evidence from a large population sample *Perspectives on Psychological Science* **8**(5):487–497, (2013).
14. Ybarra, L.M.R., Lohr, S.L.: Small area estimation when auxiliary information is measured with error. *Biometrika*, **95**(4), 919–931, (2008).