

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

Detection of spatio-temporal local structure on seismic data

Individuazione di strutture locali spazio-temporali su dati sismici

Marianna Siino, Francisco J. Rodríguez-Cortés, Jorge Mateu and Giada Adelfio

Abstract For the description of the seismicity of an area, the comparison between local features of background and induced events could be a new perspective of research. In spatio-temporal point process, local second-order statistics provide information on the relationships of each event and its nearby events. In this paper, we use a test based on local indicators of spatio-temporal association (LISTA functions) for identifying different local structures comparing the two previous sets of events. We present a simulation study on the test and show the main results of the application on Greece earthquake data.

Abstract *In questo lavoro si propone una nuova prospettiva di analisi per la descrizione della sismicità di un'area considerando il confronto delle caratteristiche locali degli eventi di fondo e quelli indotti. Nell'ambito dell'analisi dei processi puntuali di tipo spazio-temporale, le statistiche del secondo ordine locali descrivono le relazioni esistenti tra ciascun evento e i suoi più vicini. Per identificare differenze tra i due insiemi di eventi individuati in precedenza, utilizziamo un test basato sugli indicatori locali di associazione spazio-temporale, chiamati funzioni LISTA. Presentiamo uno studio di simulazione e i principali risultati applicando la metodologia sui dati sismici della Grecia.*

Key words: earthquakes; local indicators of spatio-temporal association; second-order product density function

Marianna Siino

Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Palermo, Italy

Francisco J. Rodríguez-Cortés

Department of Mathematics, Universitat Jaume I, Castellón, Spain

Jorge Mateu

Department of Mathematics, Universitat Jaume I, Castellón, Spain,

Giada Adelfio

Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Palermo, Italy, e-mail: giada.adelfio@unipa.it

1 Introduction

In an observed area, earthquake events can be considered as a realization of a marked space-time point process, where the magnitude is the mark, and a point is identified by its geographical coordinates and time of occurrence (Illian et al, 2008). Generally, the description of seismic events requires the definition of more complex models than stationary Poisson process since clustering structure characterises these events. Therefore, spatio-temporal cluster analysis has a relevant role in the comprehension of seismic processes.

Commonly, global spatio-temporal second-order summary statistics (such as the K - and pair-correlation functions) are used to detect deviations from the Poisson assumption (Gabriel and Diggle, 2009). These tools play a fundamental role in the phase of descriptive analysis, in model validation and for testing procedures giving global information of a given point pattern. An interesting question may concern if the same conclusions are valid locally, and thus, for example, in testing procedures if in subregions of the spatio-temporal window the pattern behaves differently identifying specific regions where the null hypothesis is not accepted.

Anselin (1995) proposed the idea of considering individual contributions of a global estimator as a measure of clustering under the name of Local Indicators of Spatial Association (LISA). In spatial point processes, Cressie and Collins (2001) propose a local product density function developing theoretical properties, namely first- and second-order moments, of these functions. Some applications of LISA functions are in Mateu et al (2007) and Moraga and Montes (2011). Rodríguez-Cortés (2014) and Siino et al (2016b) extend the concept of LISA function to the spatio-temporal point pattern context defining the LISTA functions. A brief summary on this methodology is in Section 2. Moreover, Siino et al (2016b) develop a testing procedure for the local structure comparing spatio-temporal point patterns based on Local indicators of spatio-temporal association (LISTA) functions described in Section 3. A simulation study is performed to illustrate that the test proposed has the prescribed size (Section 4). For the analysis in Section 5, we consider earthquakes occurred in the Hellenic area between 2005 and 2014. We aim to detect which triggered events have a significant different local cluster structure with respect to the underlying process, represented by the background events, linking the results with the geological information available in the study area.

2 Methodology

We consider a spatio-temporal point process with no multiple points as a random countable subset \mathcal{X} of $\mathbb{R}^2 \times \mathbb{R}$, where for a point $(\mathbf{u}, t) \in \mathcal{X}$, $\mathbf{u} \in \mathbb{R}^2$ is the spatial location and $t \in \mathbb{R}$ is the time of occurrence. In practice, an observed spatio-temporal pattern is a finite set $\{(\mathbf{u}_i, t_i)\}_{i=1}^n$ of distinct points within a bounded spatio-temporal region $W \times T \subset \mathbb{R}^2 \times \mathbb{R}$, where usually W is a polygon with area $|W| > 0$ and T a single closed interval with length $|T| > 0$. Considering a bounded spatio-temporal

region $A \subset W \times T$, $Y(A)$ denotes the number of the events of the process falling in A . The intensity of a process is defined as (Diggle, 2013)

$$\rho(\mathbf{u}, t) = \lim_{|\mathbf{du} \times dt| \rightarrow 0} \frac{\mathbb{E}[Y(\mathbf{du} \times dt)]}{|\mathbf{du} \times dt|}$$

where $\mathbf{du} \times dt$ is a spatio-temporal region around the point (\mathbf{u}, t) , $|\mathbf{du}|$ is the area of the spatial region, $|dt|$ is the length of the time interval and, $\mathbb{E}(Y(\mathbf{du}, dt))$ denotes the expected number of events in the infinitesimal spatio-temporal region. The process is called homogeneous or stationary when the intensity is constant, $\rho(\mathbf{u}, t) = \rho$ for all $(\mathbf{u}, t) \in W \times T$.

When the interest is in describing the spatio-temporal variability and correlations between points of a pattern, we have to consider second-order measures, such as the product density $\rho^{(2)}(\cdot, \cdot)$. This quantity provides an interpretable measure of the spatio-temporal dependence structure and it is defined as

$$\rho^{(2)}((\mathbf{u}_i, t_i); (\mathbf{u}_j, t_j)) = \lim_{|\mathbf{du}_i \times dt_i| |\mathbf{du}_j \times dt_j| \rightarrow 0} \frac{\mathbb{E}[Y(\mathbf{du}_i \times dt_i) Y(\mathbf{du}_j \times dt_j)]}{|\mathbf{du}_i \times dt_i| |\mathbf{du}_j \times dt_j|} \quad (1)$$

where $\mathbf{du}_i \times dt_i$ and $\mathbf{du}_j \times dt_j$ are small cylinders around two distinct points (\mathbf{u}_i, t_i) and (\mathbf{u}_j, t_j) .

Under the stationary case, and ignoring edge-effects, a global naive non-parametric kernel estimator for $\rho^{(2)}(r, h)$ in (1) (Rodríguez-Cortés, 2014) is given by

$$\widehat{\rho^{(2)}}_{\varepsilon, \delta}(r, h) = \frac{1}{4\pi r |B|} \sum_{i=1}^n \sum_{j \neq i} \kappa_{\varepsilon, \delta}(\|\mathbf{u}_i - \mathbf{u}_j\| - r, |t_i - t_j| - h), \quad (2)$$

where the sum is over all pairs $(\mathbf{u}_i, t_i) \neq (\mathbf{u}_j, t_j)$ of the data points, $B = W \times T$, $r > \varepsilon > 0$ and $h > \delta > 0$. The kernel function κ has a multiplicative form $\kappa_{\varepsilon, \delta}(\|\mathbf{u}_i - \mathbf{u}_j\| - r, |t_i - t_j| - h) = \kappa_{1\varepsilon}(\|\mathbf{u}_i - \mathbf{u}_j\| - r) \kappa_{2\delta}(|t_i - t_j| - h)$, where $\kappa_{1\varepsilon}$ and $\kappa_{2\delta}$ are kernel functions with bandwidths ε and δ , respectively. For an approximately unbiased edge-corrected estimator for the spatio-temporal product density see Rodríguez-Cortés (2014). The R package `stpp` (Gabriel et al, 2013) implements the main code for the computation of the the estimator in (2).

Considering the spatio-temporal product density in (1), its local version is denoted by $\rho^{(2)i}(\cdot, \cdot)$. Rodríguez-Cortés (2014) extends the operational definition of local indicator introduced by Anselin (1995), for fixed r and h , it holds that

$$\widehat{\rho^{(2)}}_{\varepsilon, \delta}(r, h) = \frac{1}{n-1} \sum_{i=1}^n \widehat{\rho^{(2)i}}_{\varepsilon, \delta}(r, h), \quad (3)$$

An unbiased edge-corrected kernel-based estimator for $\widehat{\rho^{(2)i}}(r, h)$ is given by

$$\widehat{\rho^{(2)i}}_{\varepsilon, \delta}(r, h) = \frac{n-1}{4\pi r |B|} \sum_{j \neq i} \frac{\kappa_{\varepsilon, \delta}(\|\mathbf{u}_i - \mathbf{u}_j\| - r, |t_i - t_j| - h)}{w(\mathbf{u}_i, \mathbf{u}_j) w(t_i, t_j)}, \quad (4)$$

with $r > \varepsilon > 0$, $h > \delta > 0$, for $(\mathbf{u}_i, t_i) \in W \times T$, $i = 1, \dots, n$, $w(\mathbf{u}_i, \mathbf{u}_j)$ and $w(t_i, t_j)$ are the edge-effect factors. For formal theoretical details on the LISTA functions see Rodríguez-Cortés (2014) and for the implementation the function `LISTAfunc` in the GitHub repository `pdLISTA` (Rodríguez-Cortés, 2016).

3 Testing procedure

The test procedure is an extension of the test proposed in Moraga and Montes (2011) into the spatio-temporal context. It detects differences in the local structure of two given point spatio-temporal patterns X and Z . We test the null hypothesis of no difference in the spatio-temporal local structure of X and Z with respect to the i -th point $(\mathbf{u}_i, t_i) \in X$, where the number of points in the two patterns are respectively $N(X) = n$ and $N(Z) = m$. The steps of the testing procedure are the following.

1. For each point $(\mathbf{u}_i, t_i) \in X$, for $i = 1, \dots, n$ the LISTA function $\widehat{\rho}^{(2)i}_{\varepsilon, \delta}(r, h)$ is estimated.
2. Secondly, for each fixed point $(\mathbf{u}_i, t_i) \in X$, k point patterns are generated under the null hypothesis. For each fixed point (\mathbf{u}_i, t_i) , k local spatio-temporal product density surfaces are estimated, $\widehat{\rho}^{(2)iq}_{\varepsilon, \delta}(r, h)$ for $q = 1, \dots, k$. They are summarised in terms of the average surface, denoted by $\bar{\rho}^i_{H_0}(r, h)$.
3. Based on the previous quantities, the following statistic is considered

$$T^i = \int_0^{h_0} \int_0^{r_0} \left(\widehat{\rho}^{(2)i}_{\varepsilon, \delta}(r, h) - \bar{\rho}^i_{H_0}(r, h) \right)^2 dr dh, \quad (5)$$

where r_0 and h_0 are chosen using the Diggle's rule (Diggle, 2013).

4. The theoretical distribution of our statistics under the null hypothesis is not known, so we rely on simulation-based empirical distributions. Fixing a point $(\mathbf{u}_i, t_i) \in X$, the estimated value of the statistic, is compared with the empirical distribution of the k values of $T^i_{H_0}$ with $q = 1, \dots, k$ that are obtained computing the test between the q -th generated LISTA surfaces under the null hypothesis and their sample mean function. The p -value of T^i is the following ratio $p^i = \sum_{q=1}^k \mathbf{I}(T^i_{H_0} \geq T^i) / k$. The null hypothesis is rejected if $p^i \leq \alpha$, where α is the type I error.

4 Simulation study

A simulation study with some scenarios is carried out to assess the performances in terms of type I error of the test introduced in the previous section.

The patterns are generated in the unit cube, $W \times T = [0, 1]^2 \times [0, 1]$ and varying the type of process (Poisson, Poisson cluster). We also consider $\mathbb{E}[N(W \times T)] = n + m = \{150, 300\}$ for $X \cup Z$. Under the null hypothesis, a pattern is generated with expected number of points equal to $n + m$, and the points are randomly associated to the pattern X or Z such that the number of points for the two sets is equal, and the test is computed for all the points belonging in X . For each point, the number of permutations is equal to $k = 99$.

The spatio-temporal Poisson point patterns are generated using the function `rpp` in the package `stpp` of **R**. The Poisson cluster processes are simulated using `rpcp` in `stpp`. Given the values of $n + m$ and the dispersion parameters, we control the degree of clustering by changing the expected number of parents ($n_p = \{5, 10\}$) and the number of offspring points with respect to each parent.

For each of the resulting scenarios under the null hypothesis, 100 pairs of patterns of (X, Z) are generated, and the type I error probability is defined as the proportion of points belonging to X for which the null hypothesis is rejected considering a fixed nominal value of α . Table 1 presents the average and the variance of the p -values under H_0 with the rejection rate for $\alpha = 0.05$. The statistical test exhibits acceptable empirical rejection rates for the several scenarios. There are no remarkable differences in the results when changing the intensity, the type of the process and the degree of clustering.

Scenarios		Dispersion	$n + m$	n_p	ε	δ	T_1			
X	Z	parameters					Rej.	Mean	Var	
P	P	-	150	-	0.134	0.080	0.057	0.490	0.087	
			300	-	0.111	0.069	0.055	0.487	0.086	
PC	PC	$\{h, r\} = (0.26; 0.13)$	150	5	0.094	0.069	0.053	0.500	0.086	
				15	0.114	0.074	0.048	0.493	0.084	
				300	5	0.082	0.061	0.047	0.500	0.085
				15	0.095	0.065	0.047	0.511	0.085	

Table 1: Rejection rates (Rej.) at $\alpha = 0.05$, the mean and the variance (Var.) of the p -values for the statistic. The spatio-temporal models considered are homogenous Poisson point processes (P) and Poisson cluster point processes (PC). Dispersion parameters for the PC model are given in the table, $n + m$ is the total expected number of points for $X \cup Z$, and n_p is the expected number of parents for the PC model. For each scenario, 100 simulations are considered, ε and δ are the bandwidths in space and time, respectively.

5 Application

In the seismological context, a background event refers to an earthquake that has not been triggered by another and that might be related to changes in the tectonic field. On the other hand, triggered events are thought to have been caused by a pre-

vious earthquake. Globally, these two set of events present different spatio-temporal global interaction structure, however it can be of interest to compare them focusing on a local scale.

In this application, we consider earthquake events occurred in the Greek area between 2005 and 2014 with a magnitude greater than 4 (Figure 1a), for a total number of 1105 events. Its complex spatial multiscale structure has been analysed in Siino et al (2016a).

The earthquakes are classified into background and induced events using a declustering procedure: a probability of being independent events is assigned to each one and it comes from an algorithm for the estimation of Epidemic Type Aftershocks-Sequences (ETAS) model (Ogata, 1988). We fitted the model using the R package *etasFLP* (Chiodi and Adelfio, 2014) based on the method developed in Adelfio and Chiodi (2015). We use the final probabilities provided in the last step of the iteration procedure to classify the events with a magnitude greater than 4 into the two groups, obtaining 580 background events and 525 triggered events (Figure 1a).

Considering the two clusters of independent and induced earthquakes, we would answer the following research questions: Is there a different global structure between the two point patterns? Which triggered events have a significant different local structure with respect to the underlying process (background events)? Is there any geological justification for the identified clusters?

As expected, the estimated spatio-temporal product density of the spontaneous events does not show any particular behaviour. On the other hand, for the induced seismicity, there is a spatio-temporal clustering around at $t < 300$ days and $r < 65$ kilometres, in terms of temporal and spatial distances, respectively (Figure 1b). However, we further aim to detect if we can obtain different conclusions focusing on a local scale detecting the spatio-temporal clusters.

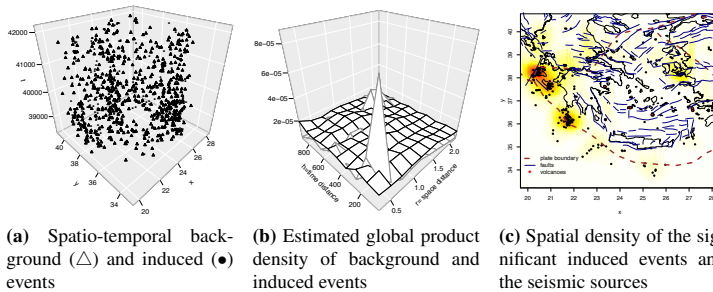


Fig. 1: (1a) Scatterplot of the spatio-temporal earthquake data classified in background events (triangles) and induced events (points) according to the procedure of declustering using the ETAS model. (1b) Estimated global product density for the background events (black surface), and induced ones (grey surface) with bandwidths $\varepsilon = 30.44$ km and $\delta = 44.14$ days. (1c) Image plot in space of the significant induced events and seismogenetic sources.

We apply the testing procedure of Section 3. The point pattern Z is represented by the background events and the events in X are the triggered ones. The representation of the results of the significant points in space allows to interpret them in relation to the geological information available in the study area (Figure 1c). We can identify some areas in which the induced events (with a magnitude greater than 4) are different in terms of spatio-temporal local structure than the background seismicity: islands of Kefalonia and Zakynthos and the Samos area (East Aegean Sea). The different behaviour is due to their specific geological characteristics and, in particular, to a higher fracturation degree of their seismogenetic volumes. These results confirm our idea that the observed seismicity is generated by a complex model, characterised by spatial-temporal interaction, with events happening at several scales, and with spatial inhomogeneity related to the geological information available in the study area.

6 Final remarks

We deal with a non-parametric testing approach for spatio-temporal point processes, in order to compare the local structure of two spatio-temporal point patterns (say X and Z). The used statistic leads to approximately valid test and the results in terms of type I error are reasonably good. Using the aforementioned test, we compare background and induced seismicity with a magnitude greater than 4 in the Greek area. It seems that the sequences of events that are strongly different to the underlying process, are placed in specific regions of the study window.

As a possible future development, we may consider further simulation scenarios to assess the power of the test. Moreover, it could be interesting to define other local tests based on the LISTA surfaces, changing what is postulated under the null hypothesis. With the analysis of the LISTA surfaces for a given point pattern, we can explore how individual points are related to their neighbouring events, clustering surfaces in order to classify points with similar spatio-temporal local structure. Moreover, we could develop a diagnostic tool based on the LISTA functions computing a weighted version of them by the inverse of the intensity function, looking for points with a more relevant contribution to the global summary statistics.

Acknowledgments

This paper has been partially supported by the national grant of the Italian Ministry of Education University and Research (MIUR) for the PRIN-2015 program (Progetti di ricerca di Rilevante Interesse Nazionale), “Prot. 20157PRZC4 - Research Project Title Complex space-time modeling and functional analysis for probabilistic forecast of seismic events. PI: Giada Adelfio”

References

- Adelfio G, Chiodi M (2015) Alternated estimation in semi-parametric space-time branching-type point processes with application to seismic catalogs. *Stochastic Environmental Research and Risk Assessment* 29(2):443–450
- Anselin L (1995) Local indicators of spatial association-lisa. *Geographical analysis* 27(2):93–115
- Chiodi M, Adelfio G (2014) etasflp: Estimation of an etas model. mixed flp (forward likelihood predictive) and ml estimation of non-parametric and parametric components of the etas model for earthquake description. R package version 10
- Cressie N, Collins LB (2001) Analysis of spatial point patterns using bundles of product density lisa functions. *Journal of Agricultural, Biological, and Environmental Statistics* 6(1):118–135
- Diggle PJ (2013) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press
- Gabriel E, Diggle PJ (2009) Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica* 63(1):43–51
- Gabriel E, Rowlingson BS, Diggle PJ (2013) stpp: An R package for plotting, simulating and analyzing Spatio-Temporal Point Patterns. *Journal of Statistical Software* 53(2):1–29
- Illian J, Penttinen A, Stoyan H, Stoyan D (2008) *Statistical Analysis and Modelling of Spatial Point Patterns*, vol 70. John Wiley & Sons
- Mateu J, Lorenzo G, Porcu E (2007) Detecting features in spatial point processes with clutter via local indicators of spatial association. *Journal of Computational and Graphical Statistics* 16(4):968–990
- Moraga P, Montes F (2011) Detection of spatial disease clusters with lisa functions. *Statistics in Medicine* 30(10):1057–1071
- Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* 83(401):9–27
- Rodríguez-Cortés FJ (2014) Modelling, estimation and applications of second-order spatio-temporal characteristics of point processes. PhD thesis, Departament de Matemàtiques; Universitat Jaume I
- Rodríguez-Cortés FJ (2016) pdLISTA: Second-order product density local indicator of spatio-temporal association function. URL <https://github.com/frajaroco/pdLISTA>
- Siino M, Adelfio G, Mateu J, Chiodi M, D'Alessandro A (2016a) Spatial pattern analysis using hybrid models: an application to the hellenic seismicity. *Stochastic Environmental Research and Risk Assessment* pp 1–16
- Siino M, Rodríguez-Cortés FJ, Mateu J, Adelfio G (2016b) An approach to hypothesis testing based on local indicators of spatio-temporal association. In: *CM-Statistics 2016*. University of Seville; Seville Spain. Blanco-Fernandez, A. and Gonzalez-Rodriguez, G.