

## 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](http://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](http://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

# 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*

# Knowledge mapping by a functional data analysis of scientific articles databases

## *Mappare la conoscenza attraverso un'analisi di dati funzionali di basi di dati di articoli scientifici*

Matilde Trevisani and Arjuna Tuzzi

**Abstract** Scientometrics studies in quantitative fashion the evolution of science focusing on the analysis of publications. One of its objectives is the development of information systems that can help to explore the enormous amount of scientific articles unceasingly published. Our study proposes an information system to reconstruct a dynamical knowledge mapping from a functional data analysis perspective. The source database is a diachronic corpus which originates a words $\times$ time-points contingency table displaying the frequencies of each keyword in the set of texts grouped by time-points in the observed time span. The information system consists of an information retrieval procedure for keywords' selection and a two-stage functional clustering to reconstruct the historical evolution of the knowledge field under investigation.

**Abstract** *La scientometria studia con un approccio quantitativo l'evoluzione della scienza attraverso l'analisi delle pubblicazioni. Uno degli obiettivi è lo sviluppo di sistemi di informazione di ausilio nell'esplorare l'enorme mole di articoli scientifici pubblicati incessantemente. Il nostro studio propone un sistema di informazione atto a ricostruire una mappatura dinamica della conoscenza secondo una prospettiva di analisi di dati funzionali. Il database di partenza è un corpus diacronico che dà origine a una tabella di contingenza parole $\times$ punti temporali contenente le frequenze di ogni parola chiave nell'insieme dei testi raggruppati per punti temporali lungo l'arco di tempo osservato. Il sistema informativo è costituito da una procedura di recupero delle informazioni per la selezione delle parole chiave e un clustering funzionale a due stadi per ricostruire l'evoluzione storica del campo di conoscenza in esame.*

---

Matilde Trevisani

Department of Economics, Business, Mathematics and Statistics, University of Trieste, Via Tigor 22, 34124 Trieste (Italy), e-mail: matilde.trevisani@deams.units.it

Arjuna Tuzzi

Department of Philosophy, Sociology, Education and Applied Psychology, Via M. Cesarotti 10/12, 35123 Padova (Italy) e-mail: arjuna.tuzzi@unipd.it

**Key words:** scientometrics, diachronic corpus, functional data analysis, cluster validation

## 1 Introduction

Scientometrics studies in quantitative fashion the evolution of science focusing on the analysis of publications. One of its major objectives is the development of information systems that can help to explore the enormous amount of scientific articles unceasingly published. The two main methods for automatically designing lexical maps are *citation-based analysis* and *co-word analysis*. Co-citation analysis maps the literature under consideration from the interaction of document citations whereas co-word analysis deals directly with the interaction of key terms shared by documents. Dynamical science mapping is another challenge that aims at describing dynamical patterns in science evolution.

In our study a dynamical knowledge mapping is reconstructed from a functional data analysis (FDA) perspective. The source database is a diachronic corpus which is a collection of texts including information on the time period to which they relate. In *bag-of-words* approaches, a diachronic corpus originates a words  $\times$  time-points contingency table displaying the frequencies of each keyword in the set of texts grouped by time-points in the observed time span. Diachronic corpora represent the ideal ground for studying the history of linguistic phenomena, e.g., when a corpus is able to reflect the relevant features of a text genre in a well-defined time period, the temporal evolution of word occurrences mirrors the historical development of the corresponding concepts [3].

This study proposes an information system consisting of (1) an information retrieval procedure for keywords' selection and (2) a functional clustering two-stage approach to identify words showing prototypical temporal patterns and cluster words portraying similar temporal patterns.

The procedure has been and is being applied to corpora of scientific papers published by leading journals of several disciplines, namely, statistics, social psychology, sociology and philosophy. This work connects to the project *Tracing the History of Words. A Portrait of a Discipline Through Analyses of Keyword Counts in Large Corpora of Scientific Literature* (University of Padova, CPDA145940, 2015-2017), involving an interdisciplinary research group whose aim is to construct chronological corpora, and, hence, to investigate whether a discipline history can be traced from analyzing the keywords' temporal pattern. Several analyses are performed to reconstruct a dynamical evolution: correspondence analysis, topic latent Dirichlet allocation, similarity analysis (using co-occurrences), and—the object of the present work—curve clustering.

## 2 Material: the corpus

Our databases are collections of articles published by a selection of premier journals of the disciplines of interest over a long time period. Text under consideration consist of titles and/or abstracts and/or full texts of the scientific papers. Time is typically discretized by years according to the cadence of volume publication.

As an example, consider one of the corpora analyzed for exploring the historical evolution of Statistics. The database is the collection of papers published by the *Journal of the American Statistical Association* (JASA, 1922-) and its predecessors, *Publications of the ASA* (1888-1912) and *Quarterly Publications of the ASA* (1912-1921). Taking into account only the texts of titles including content words and disregarding items that not refer to research papers (e.g., *List of publications*, *News*, *Comment*, *Rejoinder*), the corpus includes 10,077 titles of articles published in the period 1888-2012 (125 years, from Volume No. 1, Issue No. 1 to Volume No. 107, Issue No. 500, since at the very beginning the volumes were biennial). The corpus is composed of 87,060 word-tokens and 7,746 word-types. To solve the problem of identifying a set of keywords that prove relevant for the study of the history of Statistics, we adopt a stepwise procedure:

1. to overcome some of the limitations of analyses based on simple word-types, we replace words with stems by means of the popular Porter's stemming algorithm;
2. to take into account compounds, multi-words and sequences of words which have different meanings when they are considered in their context of use and together with adjacent words, we identify n-stem-grams;
3. to identify the most relevant statistical keywords, we match the vocabulary with popular statistics glossaries available on-line: ISI-International Statistical Institute; OECD-Organisation for Economic Cooperation and Development; Statistics.com-Institute for Statistics Education; StatSoft Inc.; University of California, Berkeley; University of Glasgow.
4. to reduce low frequency keywords we select keywords with frequencies  $\geq 10$ .

The final contingency table includes the frequencies of 900 keywords over 107 time-points.

## 3 Method: a functional clustering two-stage approach

From a FDA perspective, discrete observations  $\mathbf{y}_i = \{y_{ij}\}$  of the frequency of a keyword  $i (= 1, \dots, N)$  in the volumes  $j = 1, \dots, T$  are viewed as a realization of an underlying continuous function  $x_i(t)$ . As  $\mathbf{y}_i$  is a noisy observation of  $x_i(t)$ , an adequate model is  $\mathbf{y}_i = x_i(\mathbf{t}) + \boldsymbol{\varepsilon}_i$ , where  $\mathbf{t} = \{t_j\}$  is the finite set of time-points and  $\boldsymbol{\varepsilon}_i = \{\varepsilon_{ij}\}$  is a zero mean vector with dispersion matrix  $\Sigma_{\boldsymbol{\varepsilon}_i}$ .

For representing functional data (FD) as smooth functions one method is the basis function approach where  $x_i(t)$  is represented by a finite-dimensional linear combination  $x_i(t) = \sum_{k=1}^K c_{ik} \phi_k(t)$ ,  $c_{ik} \in \mathfrak{R}$ , for sufficiently large  $K$ , of real-valued func-

tions  $\phi_k$  called basis functions. In this study we consider B-splines as they consist in a very flexible basis for non-periodic FD. As regards the positioning of breakpoints, a direct and reasonable choice is placing knots at each time-point  $t_j$ .

We adopt the *roughness penalty* approach for estimation under which the estimate  $\hat{x}_i$  is that function minimizing the penalized residual sum of squares,  $\text{PENSSE} = \text{SSE} + \lambda \cdot \text{PEN}_r$ , where  $\text{SSE} = \{\mathbf{y}_i - x_i(\mathbf{t})\}^T W \{\mathbf{y}_i - x_i(\mathbf{t})\}$  ( $W = \Sigma_e^{-1}$ ) is the residual sum of squares,  $\text{PEN}_r = \int [D^r x_i(s)]^2 ds$  is the penalty term and  $\lambda$  is a smoothing parameter.

A standard practice for choosing  $\lambda$  is to use the generalized cross validation,  $\text{GCV}(\lambda) = T / (T - df(\lambda))^2 \text{SSE}(\hat{x}_i)$ , which provides a convenient approximation to leave-one-out CV.  $df(\lambda)$  is the effective degrees of freedom, which is monotone decreasing in  $\lambda$  with maximum equal to  $K$  when  $\lambda = 0$ .

We smooth the data by trying different spline orders combined with various roughness penalties and varying the smoothing parameter over an opportune range of values.

We adopted a distance-based approach, in particular the  $k$ -means algorithm combined with the  $L_2$  metric to measure distance between curves. Besides the  $L_2$  metric other measures of proximity can be considered, such as the  $L_1$  metric, the adaptive dissimilarity index, and the correlation-based dissimilarity [2].

Cluster validation is an essential step in the cluster analysis process. Within the approaches to cluster validation [1], the use of external information is a valuable and ultimately necessary tool. Here, external information consists of an informal assessment of subject matter experts. On the other side, a large number of indexes has been proposed in the literature for a validation based on the clustered data alone. In this study we combine a large number of internal validation indexes without integrating subject matter knowledge, so as to let the data bring out the best rated groupings. Our clustering procedure is thought of as a tool of thorough investigation before submitting the results to experts who possibly will guide towards other analyses.

## 4 Theory: corpus data transformation

The decision about what data to use is an important part of the clustering process, and often has a fundamental impact on the resulting clusters.

If we consider the keywords $\times$ time-points table by row, a typical feature of a word trajectory is a sharp peak-and-valley trend, mainly due to the sparsity affecting frequency data of a corpus. On the other hand, if we look at data by column they appear strongly asymmetrical, in particular for the marked disparity of frequency classes between the most popular words and all of the others. This is a typical feature of word-type frequency distributions aka *large number rare events* property. Lastly, the size of time-point subcorpora may vary greatly over time.

In our research, we envision several transformations which address two different objectives: whether, in assessing two curves as similar, we should consider height (word popularity) and timing (synchrony) jointly, or timing only. In the first case

we just need to normalize data by column, in the other case we need to normalize by row, or better still, since a sort of column-normalization should be regarded as preliminary, to resort to some double normalization.

The normalization step (Table 1) of our procedure provides several transformations by column ( $c_1$ - $c_5$ ) and by row ( $r_1$ - $r_5$ ).

**Table 1** Normalization plan

		normalized by column		(corpus logic)	("table" logic)	(LNRE)	
		subcorpus		column	dynamic		
normalized by row		#titles	#tokens	sum ( $\sqrt{\cdot}$ )	max. freq.	density	
Strong asymmetry	row sum	$d$	$d$	$d_1 (\chi^2)$	$d$	$d_{1b}$	$r_1$
	z-score by row	$d$	$d_2$	$d$	$d$	$d$	$r_2$
	maximum row frequency	$d$	$d_3$	$d$	$d$	$d$	$r_3$
	nonlinear transformation: $p_{x(1)}$	$d$	$d_4$	$d$	$d$	$d$	$r_4$
	nonlinear transformation: $p_{x(2)}$	$d$	$d_{4b}$	$d$	$d$	$d$	$r_{4b}$
	relative difference	$d$	$d_5$	$d$	$d$	$d$	$r_5$
		$c_1$	$c_2$	$c_3$	$c_4$	$c_5$	

Crossing a column- by a row-normalization generates a double normalization. Our comprehensive study examines all the transformations specifically indicated in the table. Here we present a small subset:  $c_2, d_1, d_3$ .

### 5 Results and conclusions

Optimal smoothing for  $c_2$  normalized data is achieved with  $m = 5$  and  $\lambda = 10^3$  ( $df = 7.7$ ) after setting a  $PEN_2$  roughness penalty, whereas for both  $d_1$  and  $d_3$  normalized data the criterion lead to  $m = 3$  and  $\lambda = 10^{1.75}$  ( $df = 7.375$ ) under a  $PEN_1$  roughness penalty. Curves are then partitioned by the  $k$ -means algorithm on the basis of the Euclidean distance. The algorithm is re-run, for each  $k$  from 2 to 26, 20 times from different initial configurations set through the  $k$ -means++ seeding method.

A set of 49 quality criteria are then computed in order to identify the best partition/number(s) of clusters. By pooling rankings from all the quality indices, the frequency of being in the top-1 up to the top-4 is calculated for each cluster number  $k$ . In general partitions into two/three clusters are the best rated. This reflects the substantial bifurcation of the historical period around the sixties at which the birth of Statistics as an autonomous and established discipline can be placed. Moreover, partitions with a number of clusters close to the maximum of the considered range have also been frequently selected. This result may be a failure due to the standard assumption of data normally distributed. From the foregoing, once discarded the solutions picking the extremes, the most selected cluster numbers are: 5 for  $c_2$ , 6 for  $d_1$  and 5 for  $d_2$ . To compare some aspects of how the three transformations affect

clustering, we consider the best partition found with the above numbers of clusters (conclusions are below).

Let us now examine some aspects of clustering, in the three cases of normalization, by varying the number of clusters (Table 2): how much groups are balanced; how many groups are singletons; how much groups are heterogeneous in being composed of words of different frequency class or popularity.

**Table 2** Balance, presence of singletons and heterogeneity of frequency classes

normalization	cluster#	cluster#											
		2	3	4	5	6	7	8	9	10	15	20	25
Quality of	$c_2$	.00	.12	.26	.29	.44	.49	.56	.59	.63	.80	.86	.91
balancing	$d_1$	.72	.93	.90	.90	.94	.96	.95	.97	.97	.98	.98	.99
	$d_3$	.84	.88	.92	.93	.93	.95	.95	.96	.97	.97	.98	.99
Number of	$c_2$	1	1	1	2	2	3	3	3	3	7	10	11
	$d_1$	0	0	0	0	0	0	0	0	0	1	1	1
singletons	$d_3$	0	0	0	0	1	0	0	1	0	3	5	5
	$c_2$	1	.50	.06	.09	.02	.02	.05	.09	.09	.11	.05	.12
Heterogeneity	$d_1$	1	1	1	.99	.99	.99	.98	.98	.97	.96	.95	.94
	$d_3$	.90	.95	.95	.93	.81	.85	.80	.82	.80	.80	.78	.77

A summary of conclusions follows.

1. Normalization by column maintains the level of word popularity differentiated and produces a dominance of high frequency words on the clustering results. Significant imbalance in cluster size, large presence of singletons, lack of heterogeneity of frequency classes in group composition and, finally, the presence of one or more “amorphous” groups, made up almost exclusively of low frequency words, are some of the most evident effects of this type of transformation.
2. Conversely, the double normalization produces groups normally well balanced both in cluster size and frequency classes, rare singletons, and almost never amorphous groups, but does lose the information on word popularity.
3. In specific, type- $d_1$  normalization is better able to recognize any group of words having “sparse” trajectories, i.e., which have experienced birth and/or death over the period considered, while the  $d_3$  variant, which more properly “normalizes” the frequency, builds the groups primarily looking at the curve shape, i.e., at if the “relative popularity” of a word has been constant over time or has fluctuated (and how) during its life cycle.

## References

1. Hennig, C., Meila, M., Murtagh, F., Rocci, R. E.: Handbook of cluster analysis. Chapman & Hall (2016).
2. Montero, P., Vilar, J.: Tslust: An R package for time series clustering, *J. Stat. Softw.* **62** (1), 1–43 (2014)
3. Trevisani, M., Tuzzi, A.: A portrait of JASA: the history of Statistics through analysis of keyword counts in an early scientific journal, *Quality and Quantity* **49** (3), 1287–1304 (2015)