



SIS 2017

Statistics and Data Science: new challenges, new generations

PROCEEDINGS OF THE CONFERENCE
OF THE ITALIAN STATISTICAL SOCIETY
28-30 June 2017 Florence (Italy)

edited by

Alessandra Petrucci

Rosanna Verde



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 Scodotto
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 Doretti, 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 Poletini, 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

Social emotional data analysis. The map of Europe

Analisi emozionale dei Social Network. La mappa dell'Europa

Felicia Pelagalli, Francesca Greco and Enrico De Santis

Abstract In this paper we present an investigation of the emotional content conveyed by words in online conversations captured on Twitter. A multivariate technique applied to co-occurrence of words together with Correspondence Analysis is adopted in order to find clusters of meaningful words detecting emotional categories that provide meaning to everyday events. Specifically, given the current historical period, where the European Union has to gain trust in its citizens, a corpus of 155000 tweets selected through the Italian keywords “Europa” and “EU” is analyzed. Results show clearly how the textual content is structured according to the different emotional expressions.

Abstract *In questo articolo è presentata un'analisi testuale che esplora il contenuto emozionale delle parole nelle conversazioni su Twitter. È stata adottata una tecnica di analisi multivariata applicata alla co-occorrenza delle parole assieme all'analisi delle corrispondenze al fine di raggruppare le parole in cluster di significato e individuare le categorie e le emozioni che danno senso agli eventi – ossia, i significati attribuiti agli eventi dagli attori partecipanti a un determinato contesto. Dato il particolare periodo storico in cui versa l'Unione Europea, che si trova a dover guadagnare la fiducia dei propri cittadini, è stato preparato ed analizzato un corpus di 155000 tweet selezionati attraverso le keyword “Europa” ed “EU”. I risultati mostrano chiaramente come il contenuto testuale è strutturato secondo le differenti espressioni emozionali del fenomeno.*

Felicia Pelagalli

Culture s.r.l., Piazza Capranica, 95 00186 ROMA, Italia,
Scuola di Specializzazione in Psicologia della Salute, Sapienza Università degli Studi di Roma, Via degli Apuli, 1 - 00185 Roma, Italia, e-mail: feliciapelagalli@yahoo.it

Francesca Greco

Dipartimento di Psicologia Dinamica e Clinica, Sapienza Università degli Studi di Roma, Piazzale Aldo Moro, 5, 00185 Roma, Italia, e-mail: francesca.greco@uniroma1.it

Enrico De Santis

Department of Computer Science, Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada, e-mail: enrico.desantis@ryerson.ca

Key words: Text Mining, Social Data Mining, Multivariate Analysis, Correspondence Analysis, Clustering.

1 Introduction

With the spread of social networks and micro-blogging platforms, statistical methodologies boosted with machine learning techniques find their natural habitat in the sea of available online data. In fact, related techniques enable us to perceive the feeling that runs through the network. An overwhelming quantity of conversations are exchanged, mostly through words in a written form. If from one side it can be possible grasping the opinions underlying the online social exchanges, from the other it is clearly interesting to have a measure of the emotional significance that gives meaning to social phenomena. Now more than ever, this knowledge can help institutions and community managers to realize people needs and problems. It is the emotion that drives us in making relation with the objects of a given context on the basis of affective symbolizations and social representations. Hence, in conveying emotions, words show the functioning structure of the mind-brain, according to a dual logic [1]: i) the asymmetrical conscious thought which allows entering in a relationship with a context or event; ii) the symmetrical emotional thinking that the context or the events immediately arouses within us. Thus, the content analysis of conversations has to catch and externalize the emotional “density” conveyed by words or chains of words, through suitable knowledge models substantiated by statistical techniques, such as the multivariate analysis. In fact, the latter, as an unsupervised technique, can find recurrences, relations between nodes of a network or can help grouping words in meaningful clusters, detecting emotional categories that provide meaning to everyday events. According to this framework, the linguistic communication can be interpreted not only on the basis of its semantic elements but also through the emotional framework that yields value to a given text. This context fits with the co-occurrence analysis of words, used as the first step of our investigation, to find associative links among words. In this study we analyze online conversations trying to discover how they are organized within the current social context and upon a given object represented by a set of keywords. Specifically, the corpus consists in 155000 tweets gathered, in the time period ranging from January 11, 2017, to February 11, 2017, through the Twitter API, filtering the stream by the Italian keywords “*Europa*” and “*UE*”. The corpus is analyzed through a pipeline of statistical and learning techniques briefly described in next section. Specifically, in order to obtain a thematic analysis based on the co-occurrence of lexical units upon the corpus at hand, a mapping of the latter in the Vector Space Model (VSM) [2] is performed. The k -means algorithm is then adopted obtaining a suitable partition through the cosine dissimilarity measure between word vectors. Finally, the Boolean contingency matrix, describing documents membership to the retrieved clusters, is analyzed with the well-known Correspondence Analysis (CA) technique. The current paper is organized as follows. In Sec. 2 we provide a brief summary

of the adopted methodology, while in Sec. 3 main results are discussed. Finally, conclusion are drawn in Sec. 4.

2 Material and methods

To finalize the herein proposed investigation, data is cleaned and pre-processed. In particular, instead of raw words, lemmas as main categories are used. Subsequently, the the most common words and the very rare words are filtered out. Lemmatization and filtering allows to obtain a more compact VSM, reducing even the sparsity of the model. We note that in the current section the formal terms “document” and “context” are interchangeable, such as “term”, “word” or “lexical units”. Following



Fig. 1 Schematic diagram of the adopted methodology for measuring the emotional structures underlying the online conversations.

the diagram of Fig. 1 the analysis presented is centered on the VSM [2], a particular vector or distributional model of meaning. VSM is based on a co-occurrence matrix, i.e. the word-document matrix, that is a way of representing how often words co-occur. From a methodological point of view the VSM embeds information retained within a corpus in a vector space representation, substantiating the distributional hypothesis according to which words that occur in similar contexts tend to have similar meanings. Lets define the term-document matrix $\mathbf{X} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_D]$ where the content of each document vector $\mathbf{d}_j = [w_1, w_2, \dots, w_V]$ is represented as a vector in the term space of dimension V that is usually the dimension of the vocabulary. A standard weighting scheme, used in the current work for w_i , is the the tf-idf (term frequency-inverse document frequency) [3], that provides higher weights to terms or words that are frequent in the current j -th document but rare overall in the collection.

In order to measure the similarity between two documents \mathbf{d}_p and \mathbf{d}_q enabling the cluster analysis, a well-suited similarity measure is used. It is the cosine similarity, that is $sim(\mathbf{d}_p, \mathbf{d}_q) = \cos(\mathbf{d}_p, \mathbf{d}_q) = \frac{\mathbf{d}_p \cdot \mathbf{d}_q}{\|\mathbf{d}_p\| \|\mathbf{d}_q\|}$.

The k -means algorithm is a *partitional* clustering algorithm [4, 5] based on squared error optimization approach. Specifically, given a set of objects (word vectors) $\mathbf{X} = \{\mathbf{d}_j\}_{j=1}^D \in \mathbb{R}^V$, where V is the dimension of data vectors, it finds a suitable partition $P = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k\}$ so that the sum of the squared distances between objects in each cluster and the respective representative element is minimized:

$$\arg \min_P \sum_{i=1}^k \sum_{\mathbf{x}_j \in \mathcal{C}_i} \|\mathbf{x}_j - \mathbf{c}_i\|^2, \quad (1)$$

where \mathbf{c}_i is the representative of the i -th cluster \mathcal{C}_i . Belonging to the family of the NP-Hard problems, a complete analytical solution is not known and k -means as greedy algorithm, can only converge to a local minimum.

CA is a statistical method useful for data visualization that is applicable to cross-tabular data such as counts, compositions or any ratio-scale data. In this work, it is performed on the Boolean contingency matrix describing the partition P [6]. Let \mathbf{P} denote a $q_r \times q_c$ data matrix with non-negative elements that sum up to 1, i.e. $\mathbf{1}_{q_r}^T \mathbf{P} \mathbf{1}_{q_c} = 1$, where in general $\mathbf{1}_q$ is a q -dimensional vector of ones and T is the transpose operator. The CA is formulated as the following least-squares problem:

$$\min_{\mathbf{A}, \mathbf{B}} \left\| \tilde{\mathbf{P}} - \mathbf{D}_r^{1/2} \mathbf{A} \mathbf{B}^T \mathbf{D}_c^{1/2} \right\|^2, \quad (2)$$

where $\tilde{\mathbf{P}} = \mathbf{D}_r^{-1/2} (\mathbf{P} - \mathbf{r} \mathbf{c}^T) \mathbf{D}_c^{-1/2}$, $\mathbf{r} = \mathbf{P} \mathbf{1}_{q_c}$, $\mathbf{c} = \mathbf{P}^T \mathbf{1}_{q_r}$, \mathbf{D}_r and \mathbf{D}_c are corresponding diagonal matrices. The column coordinate matrices \mathbf{A} and \mathbf{B} are of rank k that is the dimensionality of the approximation. By imposing $\mathbf{B}^T \mathbf{D}_c \mathbf{B} = \mathbf{I}_k$, it is possible obtaining a solution through the well-known Singular Value Decomposition: $\tilde{\mathbf{P}} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T$, where $\mathbf{\Lambda}$ is a diagonal matrix with in descending order the singular values on the leading diagonal and \mathbf{U} and \mathbf{V} are orthonormal matrices. A least-squares approximation of $\tilde{\mathbf{P}}$ is obtained by selecting the first k columns of \mathbf{U} and \mathbf{V} and the corresponding singular values in $\mathbf{\Lambda}$. Finally, the coordinate matrices are $\mathbf{A} = \mathbf{D}_r^{-1/2} \mathbf{U} \mathbf{\Lambda}$ and $\mathbf{B} = \mathbf{D}_c^{-1/2} \mathbf{V}$, so that $\mathbf{A}^T \mathbf{D}_r \mathbf{A} = \mathbf{\Lambda}^2$. Given the coordinate matrices the row coordinates are referred to as principal coordinates whereas the column coordinates are standard coordinates. The two sets of coordinates are also known as biplot and the inner-product $\mathbf{D}_r^{1/2} \mathbf{A} \mathbf{B}^T \mathbf{D}_c^{1/2}$ in (2) approximates the data. If the matrix \mathbf{P} constitutes a contingency table, $\tilde{\mathbf{P}}$ is the matrix of standardized residuals, i.e. the matrix of standardized deviations from the independence model. Hence, a low-dimensional approximation of these standardized residuals is given by the biplot coordinates in \mathbf{A} and \mathbf{B} . In other words, it can be shown that this biplot will approximate, by euclidean distances on the plot, chi-square distances in \mathbf{P} . Chi-square distance is mathematically the euclidean distance inversely weighted by the marginal totals.

3 Results

As concerns the cluster analysis the cardinality k of the partition P is set to 5. In Tab. 1 are reported the explained variances for each principal components that hereinafter are named “factors”. In Fig. 2 we can appreciate the emotional *map* of the Europe coming out from Italian tweets. It shows how discovered clusters are placed in the factorial space, whereas in Tab. 2 is reported the factors–clusters matrix that

summarizes our main findings. The emerging *map* shows on the horizontal plane a sharp contrast between the “political power” and the “populist protest”. The cluster of words \mathcal{C}_1 sees the chill and sooty European institutional places that are perceived as a remote center of power in which citizens do not definitely recognize themselves. The theme is the election of Antonio Tajani as president of the European parliament and Pittella defeat. Congratulations words, but even disappointment and irritation for who does not feel represented (*dividere, urtare, sensibilità, impera*). On the opposite side, a strong sense of helplessness regarding the big problems, such as immigrants and the economic crisis. \mathcal{C}_3 is characterized by the UE plan proposed in order to stop the sea blockade in front of Libyan territories. We have also tweets where the Italian Economy ministry is perceived as “unable”, while the former prime minister Matteo Renzi together with Angelino Alfano (current Italian foreign minister) are considered “hypocrites”. Another emerging contrast on the vertical plane is the “success of the economic power” and “people problems”. From \mathcal{C}_2 it emerges a two-speed Europe and the “economic power” represented by Germany with the chancellor Anghela Merkel and the president of the European Central Bank Mario Draghi. It is a strong power (*velocità, vincere*) that cohabits/forgets the human tragedies (*permettere, vergognarsi*). On the opposite side, \mathcal{C}_5 refers to the necessity of funds for places hit by the earthquake. Furthermore, it shows clearly the arising of new political movements, such as the one referred to Marine Le Pen in France, evidencing tension, betrayal, isolation and risks for Europe. Finally, in \mathcal{C}_4 (in a middle position on the map) we find the ambivalence fear/anguish related to the dichotomy opening–closing, where closing seems to prevail together with the fantasy of closing themselves off in the localism to avoid chaos. This is a cluster full of fears that undermine the Altiero Spinelli’s project for a united Europe. \mathcal{C}_4 is close the origin of axes on the factorial map, in fact it contains basic emotions that seem to span all the facets of the underlying discourse.

Table 1 Explained variance for each factor.

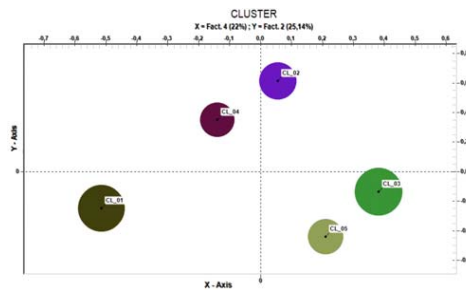
Ind	Eigenvalues	%	Cumul. %
1	0.1538	29.54	29.5438
2	0.1308	25.14	54.683
3	0.1214	23.32	77.9995
4	0.1145	22.00	100

4 Conclusion

The current paper presents an analysis of a huge corpus of tweets in Italian language based on a set of statistical techniques, specifically a Cluster analysis and a Correspondence Analysis. Unlike the current sentiment analysis techniques, the proposed

Table 2 The factor-clusters matrix.

	Factor 1	Factor 2	Factor 3	Factor 4	
\mathcal{C}_1	–	–	changing	–	Problems related to the political power unable to drive the changing.
	–	-0,2485	0,2177	-0,5145	
\mathcal{C}_2	hopes	power	changing	–	The success is related to the economic power represented by A. Merkel and M. Draghi.
	-0,5119	0,6132	0,2352	0,0558	
\mathcal{C}_3	alert	–	changing	injustice	Alert generated by the changing related to balance of powers.
	0,3088	-0,1367	0,2492	0,3808	
\mathcal{C}_5	alert	–	changing	injustice	The idea about the union with the social power of foreign countries because of the loss of identity.
	0,4653	0,3534	-0,5362	-0,1396	
\mathcal{C}_5	hopes	problems	product	–	The European genesis has a cost that causes problems: economic request for <i>help</i> and the rejection to <i>give</i> .
	-0,5718	-0,4382	-0,4911	0,2101	

**Fig. 2** The map of the Europe.

methodology takes into account the conversations on social networks like structured corpora, in which the relationships between words can be described beyond the evaluative bias (positive/negative or agree/disagree), giving rise to a dense structure of meaning. Results show clearly how the textual content is structured according to the different emotional expressions.

References

- [1] Matte Blanco. *L'inconscio come insiemi infiniti*. Biblioteca Einaudi, 2000.
- [2] Gerard Salton. The smart retrieval system – experiments in automatic document processing, 1971.
- [3] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *J. documentation*, 28(1):11–21, 1972.
- [4] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *J. ACM (JACM)*, 51(3):497–515, 2004.
- [5] Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.*, 31(8):651–666, June 2010.
- [6] M. van de Velden, A. Iodice D’Enza, and F. Palumbo. Cluster correspondence analysis. *Psychom.*, 82(1):158–185, 2017.