

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)

## **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 Mazziotto, 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<br><i>Determination of basis risk multiplier of a borrower default using survival analysis</i>                                | 1   |
| Tommaso Agasisti, Alex J. Bowers, Mara Soncin<br><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<br><i>Poverty measures to analyse the educational inequality in the OECD Countries</i>                                         | 17  |
| Mohamed-Salem Ahmed, Laurence Broze, Sophie Dabo-Niang, Zied Gharbi<br><i>Quasi-Maximum Likelihood Estimators For Functional Spatial Autoregressive Models</i>                   | 23  |
| Giacomo Aletti, Alessandra Micheletti<br><i>A clustering algorithm for multivariate big data with correlated components</i>                                                      | 31  |
| Emanuele Aliverti<br><i>A Bayesian semiparametric model for terrorist networks</i>                                                                                               | 37  |

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

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

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

|                                                                                                                                                                                     |     |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Index                                                                                                                                                                               | XI  |
| Alessandro Brunetti, Stefania Fatello, Federico Polidoro<br><i>Estimating Italian inflation using scanner data: results and perspectives</i>                                        | 211 |
| Guénael Cabanes, Younès Bennani, Rosanna Verde, Antonio Irpino<br><i>Clustering of histogram data : a topological learning approach</i>                                             | 219 |
| Renza Campagni, Lorenzo Gabrielli, Fosca Giannotti, Riccardo Guidotti, Filomena Maggino, Dino Pedreschi<br><i>Measuring Wellbeing by extracting Social Indicators from Big Data</i> | 227 |
| Maria Gabriella Campolo, Antonino Di Pino<br><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<br><i>Composite indicators for ordinal data: the impact of uncertainty</i>                                                                        | 241 |
| Stefania Capecchi, Domenico Piccolo<br><i>The distribution of Net Promoter Score in socio-economic surveys</i>                                                                      | 247 |
| Massimiliano Caporin, Francesco Poli<br><i>News, Volatility and Price Jumps</i>                                                                                                     | 253 |
| Carmela Cappelli, Rosaria Simone, Francesca di Iorio<br><i>Growing happiness: a model-based tree</i>                                                                                | 261 |
| Paolo Emilio Cardone<br><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<br><i>Comparison of conditional tests on Poisson data</i>                                                                                       | 333  |
| Riccardo D'Alberto, Meri Raggi<br><i>Non-parametric micro Statistical Matching techniques: some developments</i>                                                                           | 339  |
| Stefano De Cantis, Mauro Ferrante, Anna Maria Parroco<br><i>Measuring tourism from demand side</i>                                                                                         | 345  |
| Lucio De Capitani, Daniele De Martini<br><i>Optimal Ethical Balance for Phase III Trials Planning</i>                                                                                      | 351  |
| Claudia De Vitiis, Alessio Guandalini, Francesca Inglese, Marco D. Terribili<br><i>Sampling schemes using scanner data for the consumer price index</i>                                    | 357  |
| Ermelinda Della Valle, Elena Scardovi, Andrea Iacobucci, Edoardo Tignone<br><i>Interactive machine learning prediction for budget allocation in digital marketing scenarios</i>            | 365  |
| Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor<br><i>Nonparametric classification for directional data</i>                                                           | 371  |
| Edwin Diday<br><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<br><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

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

- Johan Koskinen, Chiara Broccatelli, Peng Wang, Garry Robins  
*Statistical analysis for partially observed multilayered networks*  
561
- Francesco Lagona  
*Copula-based segmentation of environmental time series with linear and circular components*  
569
- Alessandro Lanteri, Mauro Maggioni  
*A Multiscale Approach to Manifold Estimation*  
575
- Tiziana Laureti, Carlo Ferrante, Barbara Dramis  
*Using scanner and CPI data to estimate Italian sub-national PPPs*  
581
- Antonio Lepore  
*Graphical approximation of Best Linear Unbiased Estimators for Extreme Value Distribution Parameters*  
589
- Antonio Lepore, Biagio Palumbo, Christian Capezza  
*Monitoring ship performance via multi-way partial least-squares analysis of functional data*  
595
- Caterina Liberati, Lisa Crosato, Paolo Mariani, Biancamaria Zavanella  
*Dynamic profiling of banking customers: a pseudo-panel study*  
601
- Giovanni L. Lo Magno, Mauro Ferrante, Stefano De Cantis  
*A comparison between seasonality indices deployed in evaluating unimodal and bimodal patterns*  
607
- Rosaria Lombardo, Eric J Beh  
*Three-way Correspondence Analysis for Ordinal-Nominal Variables*  
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 |
| <p>Lucio Masserini, Matilde Bini<br/> <i>Impact of the 2008 and 2012 financial crises on the unemployment rate in Italy: an interrupted time series approach</i></p> | 673 |
| <p>Angelo Mazza, Antonio Punzo, Salvatore Ingrassia<br/> <i>An R Package for Cluster-Weighted Models</i></p>                                                         | 681 |
| <p>Antonino Mazzeo, Flora Amato<br/> <i>Methods and applications for the treatment of Big Data in strategic fields</i></p>                                           | 687 |
| <p>Letizia Mencarini, Viviana Patti, Mirko Lai, Emilio Sulis<br/> <i>Happy parents' tweets</i></p>                                                                   | 693 |
| <p>Rodolfo Metulini, Marica Manisera, Paola Zuccolotto<br/> <i>Space-Time Analysis of Movements in Basketball using Sensor Data</i></p>                              | 701 |
| <p>Giorgio E. Montanari, Marco Doretti, Francesco Bartolucci<br/> <i>An ordinal Latent Markov model for the evaluation of health care services</i></p>               | 707 |
| <p>Isabella Morlini, Maristella Scorza<br/> <i>New fuzzy composite indicators for dyslexia</i></p>                                                                   | 713 |
| <p>Fionn Murtagh<br/> <i>Big Textual Data: Lessons and Challenges for Statistics</i></p>                                                                             | 719 |
| <p>Gaetano Musella, Gennaro Punzo<br/> <i>Workers' skills and wage inequality: A time-space comparison across European Mediterranean countries</i></p>               | 731 |

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

Silvia Polettini, Serena Arima

*Accounting for measurement error in small area models: a study on generosity*  
795

Gennaro Punzo, Mariateresa Ciommi

*Structural changes in the employment composition and wage inequality: A comparison across European countries*  
801

Walter J. Radermacher

*Official Statistics 4.0 – learning from history for the challenges of the future*  
809

Fabio Rapallo

*Comparison of contingency tables under quasi-symmetry*  
821

Valentina Raponi, Cesare Robotti, Paolo Zaffaroni

*Testing Beta-Pricing Models Using Large Cross-Sections*  
827

Marco Seabra dos Reis, Biagio Palumbo, Antonio Lepore, Ricardo Rendall, Christian Capezza

*On the use of predictive methods for ship fuel consumption analysis from massive on-board operational data*  
833

Alessandra Righi, Mauro Mario Gentile

*Twitter as a Statistical Data Source: an Attempt of Profiling Italian Users Background Characteristics*  
841

Paolo Righi, Giulio Barcaroli, Natalia Golini

*Quality issues when using Big Data in Official Statistics*  
847

Emilia Rocco

*Indicators for the representativeness of survey response as well as convenience samples*  
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<br><i>On the noisy high-dimensional gene expression data analysis</i>                                                                             | 919   |
| Mirko Signorelli<br><i>Variable selection for (realistic) stochastic blockmodels</i>                                                                                                                | 927   |
| Marianna Siino, Francisco J. Rodriguez-Cortés, Jorge Mateu, Giada Adelfio<br><i>Detection of spatio-temporal local structure on seismic data</i>                                                    | 935   |
| A. Sottosanti, D. Bastieri, A. R. Brazzale<br><i>Bayesian Mixture Models for the Detection of High-Energy Astronomical Sources</i>                                                                  | 943   |
| Federico Mattia Stefanini<br><i>Causal analysis of Cell Transformation Assays</i>                                                                                                                   | 949   |
| Paola Stolfi, Mauro Bernardi, Lea Petrella<br><i>Estimation and Inference of SkewStable distributions using the Multivariate Method of Simulated Quantiles</i>                                      | 955   |
| Paola Stolfi, Mauro Bernardi, Lea Petrella<br><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<br><i>The ESSnet Big Data: Experimental Results</i> | 969   |
| Jérémie Sublime<br><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

# A comparison between seasonality indices deployed in evaluating unimodal and bimodal patterns

## *Un confronto tra indici di stagionalità utilizzati nella misura di pattern unimodali e bimodali*

Giovanni L. Lo Magno, Mauro Ferrante, Stefano De Cantis

**Abstract** This paper will discuss a recently proposed index for measuring seasonality, which is based on the solution of the well-known transportation problem. A specific characterization of the cost matrix will permit the taking into account of the cyclical structure of time periods, which characterizes the phenomenon under observation. Various features of the proposed index will be evaluated by comparing it with other indices which are commonly used in the measurement of seasonality, such as the Gini concentration index. Given the wide range of disciplines with an interest in the analysis of seasonal phenomena, the approach proposed may be of wide interest.

**Abstract** *Il presente lavoro discute un indice recentemente proposto per la misura della concentrazione stagionale e basato sulla soluzione del problema del trasporto. Una specifica caratterizzazione della matrice dei costi consente di tenere in considerazione la struttura ciclica dei periodi che caratterizza il fenomeno osservato. Diverse caratteristiche dell'indice proposto sono valutate confrontandolo con altri indici che sono comunemente utilizzati per la misura della stagionalità, come ad esempio l'indice di concentrazione di Gini. Considerata la varietà di ambiti interessati allo studio della stagionalità, l'approccio proposto può essere d'interesse da diverse prospettive.*

**Key words:** Seasonal amplitude, transportation problem, concentration index

---

Giovanni L. Lo Magno

Department of Economics, Business and Statistics, University of Palermo, e-mail: lomagno.gl@virgilio.it

Mauro Ferrante

Department of Culture and Society, University of Palermo, e-mail: mauro.ferrante@unipa.it

Stefano De Cantis

Department of Economics, Business and Statistics, University of Palermo, e-mail: stefano.decantis@unipa.it

## 1 Introduction

The notion of seasonality is very simple and very complex: referring to the former, it is a feature of many natural and human phenomena. The differing intensity of solar rays determines numerous effects on the environment, such as varying levels of humidity and temperature, but also on animal behaviour and human habits [2]. With reference to the latter, the notion of seasonality is an extremely complex concept, whose measurement and analysis is a challenge. Indeed, having reviewed the main indices used in different study contexts, we observed a lack of appropriate measures relating to seasonality, which are capable of taking the cyclical structure of time periods into account. That is, the majority of indices currently used in measuring seasonality (e.g. the Gini concentration index, the coefficient of seasonal variation or the Theil index) do not take into account the natural ordering of time periods (e.g. months). Subsequently, given a pattern in which the total amount of the phenomenon of interest is concentrated in two consecutive months (e.g. January and February) and another pattern in which the phenomenon is concentrated in two very distant months (e.g. January and August), the currently-used indices would evaluate the two patterns as the same. One of the aims of the seasonality index discussed in this paper is to overcome this issue. Thus, after a discussion of a recently-proposed index for measuring seasonality, we will evaluate some of its properties in relation to how it behaves in evaluating unimodal and bimodal distributions.

## 2 A new index for measuring seasonality and its main properties

The approach we would like to propose for measuring seasonality is based on the solution of an appropriately defined transportation problem [1]. The transportation problem is a well-known, linear, minimization problem in which the goal is to minimize the cost of transferring units from a set of warehouses to a set of customers, satisfying the constraints given by the available resources and the requested demands.

We can define a *seasonal pattern* as the vector  $\mathbf{P} = (y_1, y_2, \dots, y_T)$ , with observations for time periods from 1 to  $T$ , such that  $y_t$ , for  $t = 1, 2, \dots, T$ , is non negative. The total amount of the observed phenomenon is  $Y = \sum_{t=1}^T y_t$  and the average value is  $Y/T$ . Furthermore, let  $\mathcal{A} = \{t : y_t > Y/T\}$  be the set of time periods for which the observed value is over the average; each of these time periods has a surplus  $a_t = y_t - Y/T$ . Similarly, let  $\mathcal{B} = \{t : y_t < Y/T\}$  be the set of time periods with observed values under the average, each with deficit  $b_t = Y/T - y_t$ .

In order to eliminate seasonality, the  $T$ -dimensional pattern  $(\frac{Y}{T}, \frac{Y}{T}, \dots, \frac{Y}{T})$  can be obtained by transferring units from time periods in  $\mathcal{A}$  to time periods in  $\mathcal{B}$ . The amount which is transferred from time period  $i \in \mathcal{A}$  to time period  $j \in \mathcal{B}$  is  $x_{ij}$ . We can suppose that transferring one unit from  $i$  to  $j$  has a cost  $c_{ij}$ , thus the cost of all the transfers is  $c = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{B}} c_{ij} x_{ij}$ . The minimum cost  $c^*$  which is

required to eliminate seasonality, through the aforementioned transfers, is the cost corresponding to the optimal solution to the following transportation problem:

$$\begin{aligned}
 \min c &= \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{B}} c_{ij} x_{ij} \\
 \text{s.t.:} \\
 \sum_{j \in \mathcal{B}} x_{ij} &= a_i, \quad \forall i \in \mathcal{A} \\
 \sum_{i \in \mathcal{A}} x_{ij} &= b_j, \quad \forall j \in \mathcal{B} \\
 x_{ij} &\geq 0, \quad \forall i \in \mathcal{A}, j \in \mathcal{B}
 \end{aligned} \tag{1}$$

where the first set of constraints ensures that the amount  $a_i$  is transferred from each time period  $i \in \mathcal{A}$ ; the second set of constraints ensures that each time period  $j \in \mathcal{B}$  receives the amount  $b_j$ ; and the third set of constraints ensures that each transfer is non-negative. The minimum cost  $c^*$  corresponding to the optimal solution to the transportation problem (1) is our absolute measure for seasonality, namely

$$S(\mathbf{P}) = c^* \tag{2}$$

In [3] we demonstrated that the maximum value of  $S(\mathbf{P})$ , holding  $Y$  constant, is:

$$S_{\max}(\mathbf{P}) = \frac{Y}{T} \max_{t \in M} \left\{ \sum_{j=1}^T c_{ij} \right\} \tag{3}$$

This result permits us to construct the relative seasonality index  $S_R$  which is bounded in the interval  $[0, 1]$ :

$$S_R(\mathbf{P}) = \frac{S(\mathbf{P})}{S_{\max}(\mathbf{P})} \tag{4}$$

We recognize that the relation between time periods is cyclical. Time periods can be thought of as collocated in a circumference: the shorter arc with  $i$  and  $j$  as extremes is what is termed *cyclical distance*, and this is the unitary cost we consider for a transfer from  $i$  to  $j$ , namely:

$$c_{ij} = \begin{cases} |i - j| & \text{if } |i - j| \leq \frac{T}{2} \\ T - |i - j| & \text{otherwise} \end{cases} \tag{5}$$

Different unitary costs may be employed in the defining our seasonality index; however, when those costs defined in (5) are adopted, the  $S_R$  index has three important properties, which we deem desirable for measuring seasonality:

- *Scale invariance*:  $S_R(\mathbf{P}) = S_R(\lambda \mathbf{P})$ , with  $\lambda > 0$ .
- *Rotation invariance*:  $S_R(y_1, y_2, \dots, y_T) = S_R(y_T, y_1, y_2, \dots, y_{T-1})$
- *Sensitivity to permutations*: generally, and excluding rotations (see the “rotation invariance” property), the permutations of a pattern are evaluated in a different way (the Gini index is permutation-invariant, thus it can not capture, for example, an increase in seasonality which is consequent on a reduction in the distance between two relevant modes).

### 3 Comparing seasonality indices

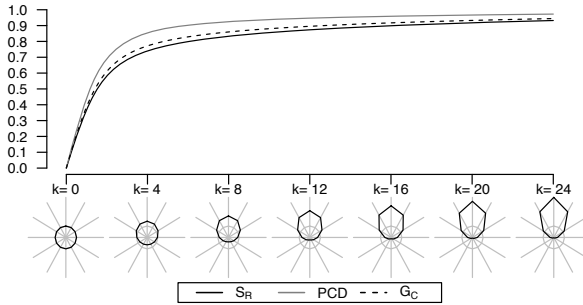
In this section we will compare: the Gini index  $G_C$  (normalized in the  $[0, 1]$  interval), the precipitation concentration index  $PCD$  [5] and the relative seasonality index  $S_R$ , when evaluating unimodal and bimodal patterns. Unimodal and bimodal patterns, obtained by adapting von Mises distributions to the discrete case, were used in these comparisons.

The von Mises distribution [4] is a circular, continuous, unimodal and symmetric distribution, with support in the  $[0, 2\pi]$  interval. Its shape depends on two parameters: the first is the expected value  $\mu$ , which corresponds to the mode; the second is  $\kappa \geq 0$  and it affects the degree of concentration. The higher is  $\kappa$ , the greater is the concentration around the mode; if  $\kappa = 0$ , then the von Mises distribution coincides with the circular uniform distribution.

In order to construct unimodal patterns of size  $T$ , we discretized von Mises distributions by partitioning their supports into  $T$  equal intervals of size  $2\pi/T$  and then calculating the probability related to each interval. Thus, the value  $y_i$  in each resulting pattern is the probability which is related to the  $i$ -th interval of the von Mises' support and, consequently,  $\sum y_i = 1$ . In this study 241 von Mises distributions were considered for the unimodal patterns; these distributions have the same  $\mu$ , but  $k = 0, 0.1, 0.2, \dots, 24$ . The parameter  $\mu$  was set to  $\pi/T$ , namely the midpoint of the first interval. Fig. 1 shows the results of all the indices which were applied to the 241 unimodal distributions under consideration. In order to clarify how the shape of the distribution changes as  $\kappa$  increases, radar charts for the distributions with  $\kappa = 0, 4, 8, 12, 16, 20, 24$  have been included in the lower part of Fig. 1. It can be observed that all the indices increase as  $\kappa$  increases, thus they correctly capture the concentration parameter  $\kappa$ . Furthermore, the relation between the three indices is  $S_R < G_C < PCD$ , with the only exception for the distribution with  $\kappa = 0$ , for which all the indices have the value 0.

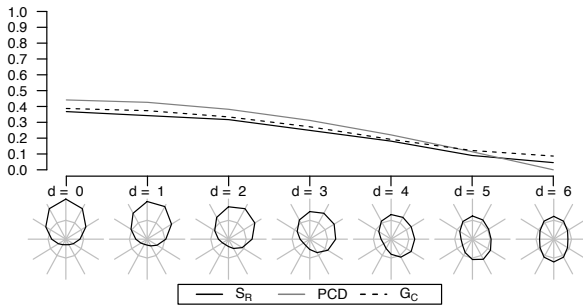
In order to observe the behaviour of the indices when they evaluate bimodal patterns, size 12 bimodal patterns were constructed as discrete versions of a 50% mixture of two von Mises distributions, both with the same  $\kappa$  parameter value. The distance  $d$  between the two modes varied between 0 (the unimodal case) and 6 (the maximum possible distance between two time periods). 1 and 5 were considered as two extreme values for  $\kappa$ , which corresponded to a low and a high concentration case around the two modes respectively. The indices values for the patterns when  $\kappa = 1$  are plotted on Fig. 2 and those for the patterns when  $\kappa = 5$  are displayed in Fig. 3.

Fig. 2 demonstrates that, for bimodal patterns where  $\kappa = 1$ , the indices decrease in value while  $d$  increases; furthermore they exhibit similar values. However, Fig. 3 also shows that the indices decrease in value while  $d$  increases, but they decrease at a quite different rate. To understand why this difference, it has to be noted that if  $\kappa$  is very high then there is a high concentration around the two modes, thus patterns where  $d > 2$  can be approximately expressed as  $\mathbf{P} = (y_1 = 0.5, 0, \dots, 0, y_{1+d} = 0.5, 0, \dots, 0)$ . Thus, for  $d > 2$ , all the patterns can be roughly considered as permutations. As the Gini index is relatively insensitive to permutations, its value is quite

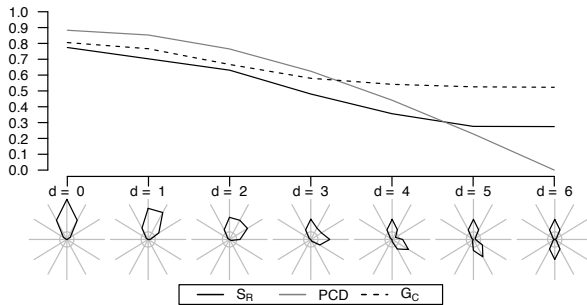


**Fig. 1** Relative seasonality index  $S_R$ , corrected Gini index  $G_C$  and precipitation concentration degree index  $PCD$ , all calculated for discretized von Mises distributions with various values of  $\kappa$  and  $T = 12$ . Radar charts are reported for the distributions with  $\kappa = 0, 4, 8, 12, 16, 20, 24$ . The regular dodecagon in each radar chart corresponds to the uniform pattern.

stable for patterns where  $d > 2$ . This is not a desirable behaviour because a seasonality index should significantly decrease when the distance between two highly concentrated modes increases. In contrast, the  $PCD$  index clearly reveals decreasing values but it is zero when  $d = 6$ . This occurs because  $PCD$  equals zero for patterns where the first half of the pattern is the same as the second, like the pattern (1, 3, 5, 1, 3, 5). A seasonality index should not return zero for patterns like these. Thus, the only index which shows a desirable behaviour is  $S_R$ , which decreases at an appreciable rate and does not reach zero when  $d = 6$ .



**Fig. 2** Relative seasonality index  $S_R$ , corrected Gini index  $G_C$  and precipitation concentration degree index  $PCD$ , all calculated for discretized, bimodal von Mises distributions with  $T = 12$ ,  $\kappa = 1$  and values of the distance between the two modes  $d = 0, 1, \dots, 6$ . Radar charts are reported for the distributions with  $\kappa = 0, 4, 8, 12, 16, 20, 24$ . The regular dodecagon in each radar chart corresponds to the uniform pattern.



**Fig. 3** Relative seasonality index  $S_R$ , corrected Gini index  $G_C$  and precipitation concentration degree index  $PCD$ , all calculated for discretized, bimodal von Mises distributions with  $T = 12$ ,  $\kappa = 5$  and values of the distance between the two modes  $d = 0, 1, \dots, 6$ . Radar charts are reported for the distributions with  $\kappa = 0, 4, 8, 12, 16, 20, 24$ . The regular dodecagon in each radar chart corresponds to the uniform pattern.

## 4 Concluding remarks

The authors of this paper contend that the challenge of measuring seasonality has not received adequate attention in the literature. For example, the Gini index does not take into account the natural ordering of time periods. In addition to permitting a specification of the cost matrix which takes into account for the cyclical structure of time periods, our seasonality index performed well when applied to unimodal and bimodal patterns. Although a set of desirable properties has been highlighted in this paper, we hope that new related issues will come to light as a result of this research. Given these challenges and considering the interdisciplinary nature of seasonal phenomena, the topic of measuring seasonality merits greater attention and from a wider range of points of view.

## References

1. Dantzig, G. B.: Linear Programming and Extensions. Princeton University Press, Princeton (1963)
2. De Cantis, S., Ferrante, M.: Seasonality and tourism. The Sage Encyclopedia of Travel and Tourism. Sage (2016)
3. Lo Magno, G.L., Ferrante, M., De Cantis, S.: A new index for measuring seasonality: a transportation cost approach. (*Submitted to Mathematical Social Sciences*).
4. Jammalamadaka, S.R., SenGupta, A.: Topics in circular statistics. Series on multivariate analysis; vol.5, World Scientific Publishing, Singapore (2001)
5. Zhang, L., Qian, Y.: Annual distribution features of precipitation in China and their interannual variations. *Acta Meteorologica Sinica* 17 (2), 146–163 (2003)