

11th Scientific Meeting of the SIS Group
"Statistics for the Evaluation and Quality in Services"

BOOK OF **SHORT PAPERS**

Editors

Andrea Bucci
Alfredo Cartone
Adelia Evangelista
Andrea Marletta



**STATISTICAL METHODS
FOR EVALUATION AND QUALITY:
TECHNIQUES, TECHNOLOGIES AND TRENDS (T³)**

EDIZIONI
Il Viandante 

**IES 2023 - Statistical Methods for Evaluation and Quality:
Techniques, Technologies and Trends (T³)**

BOOK OF SHORT PAPERS

Editors: Andrea Bucci, Alfredo Cartone, Adelia Evangelista and Andrea Marletta

Book of Short papers
11th International Conference **IES 2023**
Statistical Methods for Evaluation and Quality: Techniques, Technologies and
Trends (T³)

University 'G. d'Annunzio' of Chieti-Pescara



Scientific Committee of the group of the Italian Statistical Society on Statistics for the Evaluation and Quality of Services - SVQS

Maurizio Carpita	University of Brescia
Pietro Amenta	University of Sannio
Matilde Bini	European University of Rome
Antonello D'Ambra	University of Campania 'L. Vanvitelli'
Luigi D'Ambra	University of Naples 'Federico II'
Paolo Mariani	University of Milan 'Bicocca'
Marica Manisera	University of Brescia
Monica Palma	University of Salento
Pasquale Sarnacchiaro	University of Naples 'Federico II'

Program Committee of the conference IES 2023

Chair: Paolo Mariani	University of Milan 'Bicocca'
Matilde Bini	European University of Rome
Mario Bolzan	University of Padova
Giuseppe Bove	University of Rome 3
Eugenio Brentari	University of Brescia
Ida Camminatiello	University of Campania 'L. Vanvitelli'
Maurizio Carpita	University of Brescia
Carlo Cavicchia	Erasmus University Rotterdam
Roy Cerqueti	University of Rome 'La Sapienza'
Enrico Ciavolino	University of Salento
Corrado Crocetta	University of Bari 'Aldo Moro'
Antonello D'Ambra	University of Campania 'L. Vanvitelli'
Pierpaolo D'Urso	University of Rome 'La Sapienza'
Tonio Di Battista	University 'G. d'Annunzio' of Chieti-Pescara
Simone di Zio	University 'G. d'Annunzio' of Chieti-Pescara
Mauro Ferrante	University of Palermo
Francesca Fortuna	University of Rome 3
Michele Gallo	University of Naples 'Orientale'
Claudiu Herteliu	Bucharest University of Economic Studies
Salvatore Ingrassia	University of Catania
Ron Kenett	Samuel Neaman Institute & KPA Group
Michele La Rocca	University of Salerno
Rosaria Lombardo	University of Campania 'L. Vanvitelli'
Antonio Lucadamo	University of Sannio
Marica Manisera	University of Brescia
Andrea Marletta	University of Milan 'Bicocca'
Lucio Masserini	University of Pisa
Fabrizio Maturo	University Mercatorum
Matteo Mazziotta	ISTAT
Rodolfo Metulini	University of Bergamo
Stefania Mignani	University of Bologna
Enrico Ripamonti	University of Brescia

Silvia Salini	University of Milano
Pasquale Sarnacchiaro	University of Naples ‘Federico II’
Violetta Simonacci	University of Naples ‘Federico II’
Maurizio Vichi	University of Rome ‘La Sapienza’
Mariangela Zenga	University of Milan ‘Bicocca’

Organizing Committee

Chair: Tonio Di Battista	University ‘G. d’Annunzio’ of Chieti-Pescara
Andrea Bucci	University ‘G. d’Annunzio’ of Chieti-Pescara
Alfredo Cartone	University ‘G. d’Annunzio’ of Chieti-Pescara
Marco Di Marzio	University ‘G. d’Annunzio’ of Chieti-Pescara
Marta Di Nicola	University ‘G. d’Annunzio’ of Chieti-Pescara
Simone Di Zio	University ‘G. d’Annunzio’ of Chieti-Pescara
Stefania Fensore	University ‘G. d’Annunzio’ of Chieti-Pescara
Lara Fontanella	University ‘G. d’Annunzio’ of Chieti-Pescara
Stefano Antonio Gattone	University ‘G. d’Annunzio’ of Chieti-Pescara
Luigi Ippoliti	University ‘G. d’Annunzio’ of Chieti-Pescara
Eugenia Nissi	University ‘G. d’Annunzio’ of Chieti-Pescara
Paolo Postiglione	University ‘G. d’Annunzio’ of Chieti-Pescara
Annalina Sarra	University ‘G. d’Annunzio’ of Chieti-Pescara
Pasquale Valentini	University ‘G. d’Annunzio’ of Chieti-Pescara
Adelia Evangelista (Secretariat)	University ‘G. d’Annunzio’ of Chieti-Pescara

Editors

Andrea Bucci University of Macerata
Alfredo Cartone University 'G. d'Annunzio' of Chieti-Pescara
Adelia Evangelista University 'G. d'Annunzio' of Chieti-Pescara
Andrea Marletta University of Milan 'Bicocca'

EDIZIONI
IlViandante 

IlViandante - Copyright © 2023 Edizioni BACME S.r.l.s.
Sede legale: Via Silvino Olivieri, 111 - 66100 Chieti (CH)

ISBN 979-12-803-3369-8

DOI 10.60984/978-88-94593-36-5-IES2023

<https://doi.org/10.60984/978-88-94593-36-5-IES2023>

All rights reserved.

This work is protected by copyright law.

All rights, in particular those relating to translation, citation, reproduction in any form, to the use of illustrations, tables and the software material accompanying the radio or television broadcast, the analogue or digital recording, to publication and dissemination through the internet are reserved, even in the case of partial use. The reproduction of this work, even if partial or in digital copy, is admitted only and exclusively within the limits of the law and is subject with the authorization of the publisher. Violation of the rules involves the penalties provided for by the law.

Sponsored by



Preface

Statistical thinking, design and analysis play a crucial role in social life and are useful to society at large. Besides, promoting advanced methodological research is useful to facilitate the dissemination of ideas related to various fields of interest. For this purpose, experts in statistics, data analysis, data mining, statistical methods for decision making, machine learning and related methods come together to understand and analyse phenomena through data.

In line with this objective, the Statistics Group for the Evaluation and Quality of Services (SVQS; www.svqs.it) of the Italian Statistical Society (SIS) has been organizing the Innovation and Society (IeS) conference biennially since 2009, focusing on new developments and ideas in statistics applied to the evaluation and quality of public and private services, attracting national and international statisticians and data scientists. The meeting contributes to spot light on the main statistical approaches and methodologies for the evaluation of public services currently in use in different contexts, as well as to facilitate discussion on the impact of innovative statistical evaluation systems for these services, involving various economic and social policy actors.

The conference “Statistical Methods for Evaluation and Quality: Techniques, Technologies and Trends (T³)” recorded valuable contributions that are reported in this volume. The papers underscore how the growing availability of data has tasked social and economic actors, organizations, and researchers with the management and analysis of large volumes of unstructured and heterogeneous data. In recent years, many tools for both qualitative and quantitative models have been developed to better describe and understand complex systems and their underlying behaviors, and the papers reported in this volume bear witness to this.

Techniques, technologies and trends: the study of data complexity presents the potential to provide analyses with increased frequency and timeliness, accuracy and objectivity, and to define sustainable models. Traditional quantitative methods for capturing socioeconomic data have often shown limitations in their ability to examine underlying systems, and with the three ‘T’ just mentioned, the outlines of future developments are starting to emerge.

The volume reports 127 contributions in the following areas:

- Advanced statistical methods for pattern recognition
- Advances in statistical learning from high-dimensional data
- Data analysis for web sources
- Distance and depth-based statistical learning methods for robust data analysis

- Economics and environment
- Education and labour
- Inequalities in the labour market
- Innovations and challenges in official statistics
- Labour market: trends, perspectives and new challenges
- Methodological and applicative contributions for evaluating sustainable development
- Methodological developments and applications for the assessment of student competencies
- Networks data analysis: new perspectives and applications
- New advanced statistical methods for data science
- Recent advances in statistical learning and data analysis
- Statistical analysis and modeling of environmental pollution data
- Statistical methods and complexity for evaluation in finance
- Statistical methods and composite indicators for healthcare
- Statistical methods and models for land monitoring with spatio-temporal data
- Statistical methods for environmental monitoring and sustainability
- Statistical methods for the analysis of university student choices and academic performance
- Statistical methods for the assessment of transport services and sustainable emissions
- Statistical methods for education and educational services
- Statistics in sports
- Tourism and territory.

The Conference event attracted many contributions as well as numerous Authors, not just from Italy but also from abroad. Over the three-day meeting, the Community has the opportunity to witness some of the state-of-the arts, new trajectories, and methodological challenges in 24 solicited sessions, 7 sessions of free contributes, two round tables - organized by Maurizio Vichi and Matilde Bini respectively - and three keynotes sessions with Ron S. Kennet of Samuel Neaman Institute of Israel, Luigi D'Ambra of Federico II University of Naples, and the former Minister Enrico Giovannini from University of Tor Vergata.

Organizers

Paolo Mariani

Chair of the Program Committee

Tonio Di Battista

Chair of Local Organizing Committee

Maurizio Carpita

Coordinator of the SVQS Group

Contents

Plenary Session	1
Befitting Cross Validation with Three Case Studies	2
<i>Kenett R.S.</i>	
Solicited Session SS1 - <i>Statistical analysis and modeling of environmental pollution data</i>	3
Assessing environmental quality by clustering a structural equation model based index: An application to European cities air pollution	4
<i>Bottazzi Schenone M., Grimaccia E. and Vichi M.</i>	
Evaluating the nonlinear association between PM ₁₀ and emergency department visits	10
<i>Bucci A., Sanmarchi F., Santi L., Giostra F., Tubertini E., Rosa S., Nante N. and Golinelli D.</i>	
Estimating spatially varying Gaussian Graphical Models to unveil relationships among pollutants in the Red River Delta in Vietnam	16
<i>Pronello N., Cucco A., Ignaccolo R. and Ippoliti L.</i>	
Solicited Session SS2 - <i>Statistics in sports</i>	22
Clustering Athlete Performances in Track and Field Sports	23
<i>Argiento R., Colombi A., Modotti L. and Montagna S.</i>	
A Cross-Country Analysis of Engagement in Physical Activity and Sport Practice Learnt from Eurobarometer Survey Data	29
<i>Simone R.</i>	
Strong eras in male professional tennis	35
<i>Breznik K., Candila V., Milekhina A. and Restaino M.</i>	
NonParametric Combination method for data analytics in basketball matches	41
<i>Barzizza E., Biasetton N., Ceccato R., Disegna M. and Vezzosi G.</i>	
Solicited Session SS3 - <i>Statistical methods for the analysis of university student choices and academic performance</i>	47
The influence of labor market conditions on students' career disruption: first insights from Italy	48
<i>Usala C., Sulis I. and Porcu M.</i>	
Socio-economic aspects that may affect South-North students' mobility in Italy	53
<i>Genova V.G. and Boscaino G.</i>	
An analysis of student's performance in bachelor's degree	59

<i>La Rocca M., Niglio M. and Restaino M.</i>	
An exploratory strategy for analyzing students' mobility data	65
<i>Primerano I. and Giordano G.</i>	
Solicited Session SS4 - <i>Statistics in football</i>	71
Community Detection in Sport Market Networks: The Case of Italian Professional Football	72
<i>Rondinelli R. and Ievoli R.</i>	
An Original Application to Football of PLS-SEM for the xG Model	78
<i>Cefis M. and Carpita M.</i>	
Performance Assessment of Football Players and Combined Permutation Tests with application to Home-Field Advantage	84
<i>Bonnini S.</i>	
A First Proposal of the Triad Census for Weighted Networks: an Application to Football	90
<i>Rondinelli R. and Palazzo L.</i>	
Solicited Session SS5 - <i>Advances in statistical learning from high-dimensional data</i>	96
PCA approaches for vector functional time series	97
<i>Aguilera A.M., Alonso F.J. and Acal C.</i>	
Conformal Prediction for Functional Kriging Models	102
<i>Diana A., Romano E. and Adzic J.</i>	
Measuring Public-Private Connectedness in Financial Markets	108
<i>Sánchez García J. and Cruz Rambaud S.</i>	
An original approach to anomalies in intertemporal choices through Functional Data Analysis: Theory and application for the study of Hikikomori syndrome	112
<i>Martino R., Ventre V., Cruz Rambaud S. and Maturo F.</i>	
Solicited Session SS6 - <i>Labour market: trends, perspectives and new challenges</i>	118
Enriching Job Vacancy Official Information with Online Job Advertisements: Chances and Limits	119
<i>Lucarelli A. and Righi A.</i>	
Innovation in Management: towards the Open Manager	126
<i>Bruttini P., Gallo M., Mariani P. and Menini T.</i>	
Solicited Session SS7 - <i>Data analysis for web sources</i>	132
Enhancing SMEs default prediction with web-scraped data	133
<i>Crosato L., Domenech J. and Liberati C.</i>	
Web data as enabler for informed decisions in Labour Market	137
<i>Maggioni G.</i>	
The metaverse & luxury fashion brands: strategic communication exercise	141
<i>Forciniti A. and Zavarrone E.</i>	
Increasing the Geographical Granularity of Economic Indicators with Google Trends	147

Domenech J. and Marletta A.

Solicited Session SS8 - <i>Methodological and applicative contributions for evaluating sustainable development</i>	153
Evaluating sustainable development in EU countries through synthetic indicators	154
<i>Alaimo L.S. and Cucci M.</i>	
Naples and tourism sustainability: A survey of citizens' perceptions	160
<i>Aria M., Pagliara F., D'Aniello L. and Della Corte V.</i>	
Modelling inequalities for sustainable development in Italy countries	166
<i>Musella M., Borrata G., Camminatiello I. and Lombardo R.</i>	
Food Security and Sustainability: A Science Mapping Analysis	172
<i>Piscitelli A.</i>	
Solicited Session SS9 - <i>Inequalities in the labour market</i>	178
Skill similarities across Italian regions: an analysis based on the online job advertisements	179
<i>Kahlawi A., Buzzigoli L., Grassini L., Martelli C. and Giambona F.</i>	
Italian Labour Market reform and gender inequalities	185
<i>Marini C. and Nicolardi V.</i>	
Intergenerational transmission of disadvantages in the Italian labour market: evidence from AD-SILC data	191
<i>Busetta A., Fabrizi E., Ragozini G. and Sulis I.</i>	
Solicited Session SS10 - <i>Statistical methods and complexity for evaluation in finance</i>	197
Financial networks resilience and shocks propagation	198
<i>Cerqueti R., Cinelli M., Ferraro G. and Iovanella A.</i>	
How the choice of one parameter impacts the numerical stability of the efficient frontier	204
<i>Fassino C. and Uberti P.</i>	
Dynamic shrinkage for minimum variance combination of forecasts	210
<i>Mattera R.</i>	
Exploring the perception of the gender issue of Italian female entrepreneurs	217
<i>Castellano R., Riccioni J. and Rinaldi A.</i>	
Solicited Session SS11 - <i>Networks data analysis: new perspectives and applications</i>	223
Describing Italian mobility trajectories in higher education	224
<i>Genova V.G., Giordano G., Ragozini G. and Vitale M.P.</i>	
Collaboration networks: methodological issues and updated empirical evidence on Italian statisticians	230
<i>De Stefano D., Fabbrucci Barbagli A.G., Santelli F. and Zaccarin S.</i>	
Mapping Ashtma Complexity with Graph Theory: an Integrative Approach	236
<i>Cucco A., Simpson A., Murray C., Fontanella S. and Custovic A.</i>	
Investigating the patterns of Italian internal mobility: a network analysis at provincial level	242

Sarra A., D'Ingiullo D., Evangelista A., Nissi E., Quaglione D. and Di Battista T.

Solicited Session SS12 - <i>Innovations and challenges in official statistics</i>	248
Formal and informal networks of care for the elderly: regional profiles compared	249
<i>Sicuro L., Tucci D. and Coniglio R.</i>	
Gender Gap: a multidimensional approach	255
<i>Acampora C., Fusco D., Liguori M.A. and Pagliuca M.M.</i>	
Using Whatsapp in Official Statistics: a New tool for managing the Agriculture Census	259
<i>Fabi C.</i>	
Solicited Session SS13 - <i>Statistical methods and composite indicators for healthcare</i>	265
Longitudinal composite indicators to measure the quality of health services	266
<i>Crocetta C., Antonucci L., Cataldo R. and Mazza R.</i>	
Past and Future of Doctor-Patient Communication	272
<i>Tedesco N., Zavarrone E. and Forciniti A.</i>	
Network Analysis approach to customer satisfaction and service quality detection: an application to health-care services	277
<i>Crocetta C., Grassia M.G., Marino M., Mazza R., Simonacci V. and Stavolo A.</i>	
A project evaluation study on multiset Likert scale data	283
<i>Simonacci V., Marino M., Grassia M.G. and Gallo M.</i>	
Solicited Session SS14 - <i>Distance and depth-based statistical learning methods for robust data analysis</i>	289
Robust distance-based predictive models	290
<i>Boj E., Grané A. and Parron D.</i>	
Data depth for mixed-type data through multidimensional scaling. An application to biological age imputation	294
<i>Cascos I., Grané A. and Qian J.</i>	
A compared protocol to improve clustering procedures	298
<i>Grané A., Riani M. and Salini S.</i>	
Robust diagnostics for Linear Mixed Models with the Forward Search . . .	304
<i>Corbellini A., Grossi L. and Laurini F.</i>	
Solicited Session SS15 - <i>Advanced statistical methods for pattern recognition</i>	310
Unsupervised classification of NPLs recovery curves	311
<i>Carleo A. and Rocci R.</i>	
Living alone in Italian municipalities	317
<i>Vellucci P., Benassi F., Naccarato A. and Gallo G.</i>	
Supervised learning from high-dimensional data through dynamic updating of functional classification rules	323

<i>Maturo F., Fortuna F. and Di Battista T.</i>	
Assessing the effectiveness of coordination among public authorities in cohesion expenditure	329
<i>Coco G., Monturano G. and Resce G.</i>	

Solicited Session SS16 - *Recent advances in statistical learning and data analysis* 335

A Predictive Functional Principal Component Analysis of Well-Being Data	336
<i>Marcis L., Pagliarella M.C. and Salvatore R.</i>	
Detecting the partition in the extended hierarchy of a dendrogram: an application on biomedical data	343
<i>Policastro V., Palazzo L. and Vistocco D.</i>	
Concordance measure for rankings	350
<i>Bissiri P.G. and Nai Ruscone, M.</i>	
Quadratic discriminant scoring for selecting clustering solutions	355
<i>Coraggio L. and Coretto P.</i>	

Solicited Session SS17 - *Statistical methods for education and educational services* 361

Association between INVALSI scores and students' mobility in Italy: a preliminary analysis	362
<i>Bacci S., Bertaccini B., Lombardi G. and Tocchioni V.</i>	
Modelling Responses and Response Times: an application to Mathematics INVALSI data	368
<i>Bungaro L., Desiderati R. and Mignani S.</i>	
Latent potential outcomes: An analysis of the effects of programs aimed at improving students' non-cognitive skills	374
<i>Pennoni F., Bartolucci F. and Vittadini G.</i>	
Cognitive Skills and Non Cognitive Skills to Analyze School and Students Performances	380
<i>Vittadini G.</i>	

Solicited Session SS18 - *Statistical methods for the assessment of transport services and sustainable emissions* 386

Sustainability assessment of urban transport by an LCA comparison on different technologies vehicles	387
<i>Della Ragione L. and Meccariello G.</i>	
Passenger comfort prediction via time-series classification	393
<i>Vanacore A., Pellegrino M.S. and Ciardiello A.</i>	
A statistical model to analyse driving behavior: a case study	399
<i>Rodia G., Sarnacchiaro P. and Acciarino V.</i>	
Aggregating judgments in non negotiable group decisions in transport system	403
<i>Amenta P. and Lucadamo A.</i>	

Solicited Session SS19 - *New advanced statistical methods for data science* 408

A unified framework for two-dimensional clustering on preference-approvals: an analysis of Eurobarometer data	409
<i>Albano A., Sciandra M. and Plaia A.</i>	
Pandemic Data Quality Modelling: A Bayesian Approach	415
<i>Ferrari L., Manzi G., Micheletti A., Nicolussi F. and Salini S.</i>	
Explainable AI for Peer-to-Peer Credit Risk Management	421
<i>Babaei G., Pagnottoni P. and Do T. T.</i>	
Tackling misclassification in surveys about undeclared work via the EM algorithm	427
<i>Arezzo M.F., Guagnano G. and Vitale D.</i>	

Solicited Session SS20 - *Tourism, territory and data analysis* 432

Tourism, sustainability, and territorial impact: an input-output analysis . .	433
<i>Garau G. and El Meligi A.K.</i>	
The Impact of Big Data in Tourism	439
<i>Ciuffreda R., Choedon C. and Simonetti B.</i>	
Neural network-based prediction of domestic tourists' length of stay in Italy	443
<i>Antolini F. and Cesarini S.</i>	
The management of cultural heritage in contexts of undertourism: a model for assessing the economic sustainability of public-private partnerships	450
<i>Calabrò F.</i>	

Solicited Session SS21 - *Statistical methods and models for land monitoring with spatio-temporal data* 457

Geo-referenced data and complex networks for measuring road accident risk	458
<i>Cantaluppi G., Clemente C., Della Corte F. and Zappa D.</i>	
A comparison of geospatial models for car crash risk	464
<i>Cantaluppi G., Giardino G. and Zappa D.</i>	
Geostatistical modelling of livestock-related PM _{2.5} pollution and scenario analysis for policymakers - Work in progress	470
<i>Fassò A., Rodeschini J., Fusta Moro A. and Finazzi F.</i>	
Functional clustering methods for space-time big data from mobile phone networks	476
<i>Perazzini S., Metulini R. and Carpita M.</i>	

Solicited Session SS22 - *Methodological developments and applications for the assessment of student competencies* 483

Modeling the main drivers of mathematical literacy of school-leaving stu- dents. Some evidence from the Invalsi tests	484
<i>Davino C., Palumbo F., Romano R. and Vistocco D.</i>	
Educational Data Mining: clustering students' performance over time . . .	490
<i>Taraborrelli G. and Farnè M.</i>	
The nexus of cultural capital with participation in early childhood education	496
<i>Ripamonti E.</i>	
High- and Low-Performing students and future career: a gender and social issue	502
<i>Falzetti P. and Ricci R.</i>	

Solicited Session SS23 - <i>Statistical methods for environmental monitoring and sustainability</i>	508
Clustering spatial data through optimal transports	509
<i>Balzanella A. and Verde R.</i>	
New interpretative insights for environmental air quality by means of FDA	514
<i>Terzi S., Naccarato A. and Fortuna F.</i>	
A Bayesian State-Space Model to Mitigate Unmeasured Confounding . . .	520
<i>Zaccardi C., Valentini P. and Ippoliti L.</i>	
Mining social media data for damage assessment in environmental disasters	526
<i>del Gobbo E., Cafarelli B., Ippoliti L. and Fontanella L.</i>	
Solicited Session SS24 - <i>Satisfaction and behavior in tourism</i>	531
The evaluation of the hotel stay through a new development of correspondence analysis coping with ordinal variables	532
<i>D'Ambra A. and Amenta P.</i>	
Assessing the role of knowledge and authenticity in the formation of attendee loyalty at cultural festivals	536
<i>Rivetti F., Lucadamo A. and Rossi C.</i>	
Residents' Opinions and Perceptions of Tourism Development in the Historic City of Matera	542
<i>Sarnacchiaro P., Di Gennaro R. and Di Taranto E.</i>	
Exploring tourism at religious sites: The case of Assisi	547
<i>Rivetti F., Dini M. and Splendiani S.</i>	
Session of free contributes SFC1 - <i>Education and labour</i>	552
Local concordance among the items of questionnaires on student's opinion (OPIS)	553
<i>Terzi S. and Petrarca F.</i>	
High School Proficiency of Future University Students: An Analysis based on INVALSI Data	560
<i>Santelli F., Di Credico G. and Di Caterina C.</i>	
Employment vulnerability of immigrants in the labour market – Does origin matter?	566
<i>Bittaye M.</i>	
The effect of pricing policies on students' use of university canteens	572
<i>Masserini L., Bini M. and Lorenzoni V.</i>	
Gig workers' identikit	576
<i>Zavarrone E. and Forciniti A.</i>	
Session of free contributes SFC2 - <i>Tourism and territory</i>	582
Multi-source approach for statistics in tourism sector	583
<i>Bianchino A., Fusco D., Giordano P., Liguori M.A. and Summa D.</i>	
Statistical analysis of tourism sustainability in Campania: post Covid-19 review	589
<i>Giactalone M., Basile V. and Bellucci M.</i>	
Investigating recent changes in dietary behavior	596
<i>D'Uggento A., d'Ovidio F.D., Toma E. and Onorati M.G.</i>	

Depopulation in the Abruzzo municipalities	603
<i>Carulli A.L., Di Spalatro D. and Valentini A.</i>	
The Productions System of Inland Areas	607
<i>Madia Carucci A.M. and Regano A.</i>	
Session of free contributes SFC3 - <i>Social issues</i>	613
Modelling the gender gap in youth mortality with an Age-Period-Cohort analysis	614
<i>Lanfruti Baldi G. and Nigri A.</i>	
Random forest for classifying odor emission sources	620
<i>Distefano V., Palma M., De Iaco S. and Mazuruse G.</i>	
An Experimental Annotation Task Investigating Annotator Agreement Within a Misogynistic Dictionary and Corpus	626
<i>Tontodimamma A., Ignazzi E., Anzani S., Fontanella L. and Di Zio S.</i>	
Statistical analysis of COVID19 impact on Italian mortality	632
<i>Franchetti G. and Politano M.</i>	
Measuring multidimensional deprivation using objective and subjective data: an application of the Voronoi ranking method	638
<i>Ciommi M., Mariani F., Polinesi G. and Recchioni M.C.</i>	
Session of free contributes SFC4 - <i>Trends</i>	644
The role of big data analytics in circular supply chains: A bibliometric analysis	645
<i>Khan F. and Rapposelli A.</i>	
Estimation of the ranking of incentive policies for the adoption of 4.0 technologies	652
<i>Bonnini S. and Borghesi M.</i>	
Risk Management and Future Scenarios. A proposal based on a mixed-method approach	658
<i>Di Zio S., Bolzan M., Marozzi M. and Scioni M.</i>	
Explainable artificial intelligence (XAI) through artificial intelligence from a human in the loop (HITL) perspective: an interview with ChatGPT	664
<i>Santarcangelo V., Lamacchia A., Vitullo S., Di Lecce M. and Giacalone M.</i>	
Relevance in official statistics: information needs, satisfaction with data quality, some results and future perspectives	670
<i>Loporcaro M.F.</i>	
Session of free contributes SFC5 - <i>Economic issues</i>	676
Evolutionary trends of start-ups in Italy: a case study	677
<i>Duttilo P., Caruso G., Iannone B. and Gattone S.A.</i>	
Permanent establishments and efficiency analysis with global enterprises	682
<i>Frenda A. and Sepe E.</i>	
Techniques and constructs in some recent market and organizational research	685
<i>Sciascia I.A.</i>	
The value of buildings in the Italian general government balance sheet	689
<i>Santoro P. and Regano A.</i>	

Linear and nonlinear factors affecting default risk in the peer-to-peer lending market	695
<i>Giordano F., Milito S. and Parrella M.L.</i>	
Session of free contributes SFC6 - <i>Methodological issues</i>	701
Macroeconomic Time Series Classification by Nonparametric Trend Estimation	702
<i>Feo G., Giordano F., Niglio M. and Parrella M.L.</i>	
A Normalization Method for Space-time Analysis of Evaluation and Quality Indicators	709
<i>Mazziotta M. and Pareto A.</i>	
Unveiling Latent Structures: exploring Multidimensional IRT Models using Dirichlet Process Mixtures	713
<i>Valentini P., Fontanella S. and Fontanella L.</i>	
On a technique to detect accounting data manipulation	718
<i>Passamonti C.</i>	
Session of free contributes SFC7 - <i>Economics and environment</i>	724
Determinants of Water Conservation Behaviour and Spatial Heterogeneity in their Coefficients	725
<i>Mammadli R. and Gigliarano C.</i>	
Modeling the economic burden of grass pollen allergoid immunotherapy . .	731
<i>Bilancia M. and Di Bona D.</i>	
Swine fever in Liguria: who does pay for economic losses? A causal analysis	735
<i>Baggetta C., Cavalletti B. and Corsi M.</i>	
Measuring Multidimensional Poverty of the Italian Regions in the era of COVID-19	741
<i>Chelli F.M.C., Ciommi M., Gigliarano C. and Polinesi G.</i>	
Drivers of inflation: relationships changing over time	747
<i>Sokolenko O., Palumbo A., Fortuna F., Naccarato A. and Marie J.</i>	
Environmental accounting and sustainable cities: an explorative bibliometric-based literature analysis	753
<i>Kaur A., Javed A., D'Andreamatteo A. and Rapposelli A.</i>	

Plenary Session

1. *Befitting Cross Validation with Three Case Studies* (Kenett R.S.)

A Predictive Functional Principal Component Analysis of Well-Being Data

Una Analisi Predittiva delle Componenti Principali Funzionali dei Dati sul Benessere

Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore

Abstract We consider the case of the principal components of a multivariate random vector that obeys a linear mixed model. The random vector itself then lies in a lower dimensional subspace. This situation suggests that this subspace can be modeled by the probabilistic (random-effects) principal components. We employ a linear predictor adjusted by the residual part of the probabilistic principal components, that results not explained by the longitudinal time-varying linear model. A new predictor is given, employing the vector of scores that comes from that principal components, and the resulting scores approximated by spline curves as functionals. The application to the official Italian well-being data shows some of the features of the method.

Abstract *In questo articolo, consideriamo il caso delle componenti principali di un vettore casuale multivariato che segue un modello misto lineare. Il vettore casuale stesso risiede, quindi, in un sottospazio di dimensione inferiore. Questo sottospazio può essere modellato dalle componenti principali probabilistiche (ad effetti casuali). Facciamo uso di un predittore lineare corretto dalla parte residuale dei componenti principali probabilistici, che risulta non spiegato dal modello longitudinale lineare time-varying. Viene dato un nuovo predittore, utilizzando il vettore dei punteggi che proviene da quei componenti principali, e i punteggi risultanti approssimati da curve spline. L'applicazione ai dati ufficiali italiani sul benessere mostra alcune delle caratteristiche del metodo.*

Key words: multivariate random vector, probabilistic principal components, functional principal components, linear mixed model, well-being

Laura Marcis

Department of Economics and Law, University of Cassino and Southern Lazio; e-mail: laura.marcis@unicas.it

Maria Chiara Pagliarella

Department of Economics and Law, University of Cassino and Southern Lazio; e-mail: mc.pagliarella@unicas.it

Renato Salvatore

Department of Economics and Law, University of Cassino and Southern Lazio; e-mail: rsalvatore@unicas.it

1 Introduction

Principal component analysis (PCA) is recognized as one of the most employed methods to reduce dimensionality, by means of projection of a set of variables in a subspace of them. By summarizing and allowing to visualize data, and, at the same time, minimizing the loss of information in the lower dimensional space, in many cases principal components (PCs) lead to a better assessment of the bundled statistical information, seized by the original variables [2]. In the case of studying data with measurements over time of a multivariate random vector, the goal is to assume that the principal components are curves considered as functionals [4]. These functionals change with a continue variable, that is time. Then, functional principal component analysis is a standard PCA performed by sampling units curves at each time instant. The evaluation of the functionals are related to smoothing time PC points, and the resulting curves considered as the principal sources of variability. This method involves an integral transform, which reflects the functional nature of the data, whose analytic solutions cannot, in general, be determined.

Another important and recurrent way of approaching redundancy and, in general, the recursive informative content of multivariate sample data, is given by considering a common subset of covariates which the population obeys. Two main cases in the literature are deemed representative of the joint dependence on a multivariate vector, the PCs “with covariates” or Partial PCA, and the redundancy analysis. Although these last represent a very useful tool in some cases, the deployment of linear models to explain part of the sample variability has had a remarkable development in the last years. One of these studies brings into play the role of prediction by linear statistical models.

Tipping and Bishop [7] had already introduced the notion of prediction for the PCs. They called “Probabilistic PCA” (PPCA) the model behind the PCA, which parameters are estimated with the Expectation-Maximization algorithm. Instead of a “fixed effects PCs”, as the traditional linear regression PCA model, the PPC are random variables. This condition suggests, on the one hand, the Bayesian approach to handle the estimates for the PPC linear model and, on the other hand, to predict PCs under its meaning within random linear models theory [5]. Given normality of the error $\varepsilon \sim N(0, \sigma^2 I)$, for a linear model $\tau = B\lambda + \varepsilon$ - in case of the vector λ random - the likelihood is based on the conditional distribution $\lambda|\tau \sim N[E(\lambda|\tau), var(\lambda|\tau)]$. Moreover, it is known [6] that $E(\lambda|\tau) = \tilde{\lambda}$ is the “Best Prediction” (BP) estimate, with $var(\tilde{\lambda} - \lambda) = E_{\tau}[var(\lambda|\tau)]$. Therefore, given a linear mixed model (LMM, [1]) for τ , with $E(\tau|\lambda) = \lambda$, the model parameters are the realizations of random variables. Thus, given the BP estimates of the PPCs λ , $\tilde{\lambda} = E(\lambda|\tau)$, the vector $\tilde{\tau} = B\tilde{\lambda}$ represents the BP of the p -variate vector.

In the present paper we introduce a model of PPCs combined with a multivariate longitudinal linear predictor, considering then the random vector effectively represented by the subspace of the resulting “shifted” PPCs (SPCs). The new smoothed SPCs can be considered as time-varying functional principal components, because they combine the linear model and the PPC’s, and carry out simultaneously the lin-

ear predictor and the contribution given by the PPCs not “explained” by the linear predictor itself.

An application to the official Well-being Italian indicators shows some of the features of the method.

2 Theory

In the sequel we report the following symbols, giving the model specification. We consider n as the number of subjects in the longitudinal LMM ($i = 1, \dots, n$), $N = n \times T$ as the total number of observations considered, with $t = 1, \dots, T$ the time instants. Then p ($j = 1, \dots, p$) is the number of the response dependent variables, l the number of the linear model covariates, and s is the dimension of the effective PC subspace. Consider Θ as the $N \times p$ sample matrix of the p -variate $p \times 1$ random vector θ , with N as the total number of the units given by the sample. Moreover, consider that the vector θ obeys the linear model:

$$\theta = \beta'x + u', \quad (1)$$

where x the $l \times 1$ vector of covariates, β the $l \times p$ matrix of the regression effects, u is the vector of the p -variate random effect, with $u \sim N(0, \Sigma_u)$, $\Sigma_u = cov(u)$. Furthermore, we consider at the same time that the multivariate random vector θ obeys the following linear model:

$$\theta = Ab + \varepsilon, \quad (2)$$

in which A is $p \times s$ a loading matrix of eigenvectors, b is the random vector of PPCs, and ε is a vector of isotropic error, with $\theta \sim N(\mu, A\Psi A' + \sigma_e^2 I)$, $b \sim N(0, \Psi)$, $\Psi = diag(\psi_1, \dots, \psi_s)$, $s < p$, and $\varepsilon \sim N(0, \sigma_e^2 I)$. When a sample of N observations is given, an $N \times p$ matrix Y of observations from the random vector θ is simply modeled as $Y = \Theta + E$, with the “sampling error” $Np \times Np$ covariance matrix $cov(vec(E)) = (\Sigma_e) \otimes I_N$ (\otimes is the Kronecker product), $e \sim N(0, \Sigma_e)$, $\Sigma_e = var(e)$. Thus, models 1 and 2 are rewritten as $Y = \Theta + E = X\beta + ZU + E$, with the 1 that becomes $\theta = \beta'x + u' + e'$, and $Y = \Theta + E = BA' + \Xi + E = BA' + \Gamma$, with the 2 that is $\theta = Ab + \varepsilon + e = Ab + \gamma$, respectively. The model errors u , ε , and e , are mutually independent. The matrix Z represents the $N \times n$ design matrix of random effects and E is the $N \times p$ matrix of the residual errors of the multivariate LMM, B is the $N \times s$ matrix of the PPCs that lie in the s -dimensional subspace, Ξ is the $N \times p$ matrix of the isotropic errors of the model 2. The models 1 and 2 have the following conditional expectations and variances:

$$\begin{aligned} E(\theta|y) &= \tilde{\theta}_y = y - E(e|y) = y - cov(e, y)var(y)^{-1}y = y - var(e)Py, \\ var(\theta|y) &= var(\theta) - cov(\theta, y)var(y)^{-1}cov(y, \theta) = var(e) - var(e)Pvar(e), \end{aligned}$$

for the model in 1, where $P = \Sigma_y^{-1}(I - P_X)$, $\Sigma_y = var(y)$, and P_X is the projection matrix. For the model in 2:

$$\begin{aligned}
 E(b|\theta) &= E(b) + \text{cov}(b, \theta)\text{var}(\theta)^{-1}(\theta - \mu) \\
 &= \text{cov}(b, \mu + Ab + \varepsilon)C^{-1}(\theta - \mu) = \Psi A' C^{-1}(\theta - \mu), \\
 \text{var}(b|\theta) &= \text{var}(b) - \text{cov}(b, \theta)\text{var}(\theta)^{-1}[\text{cov}(b, \theta)]' \\
 &= \Psi - \Psi A' C^{-1} A \Psi, \quad C = A \Psi A' + \sigma_\varepsilon^2 I.
 \end{aligned}$$

Based on some results on linear projections, i.e., given the random variable y , and the $1 \times j, 1 \times k$ random vectors x, z , with positive definite covariance matrix of $(y, x, z)'$, then we get for the linear projection $L(y|x, z) = L(y|x) + [z - L(z|x)]\gamma$, where $\gamma = \text{var}(z|x)^{-1}[\text{cov}(y, z|x)]'$. We have the following:

Proposition 1. *Given the model 2 for the p -dimensional random vector θ , with $b = \bar{F}'(\theta - \varepsilon)$, and under the models in 1 and 2, the multivariate Best Predictor based on (y, b) , $E(\theta|y, b)$, is:*

$$E(\theta|y, b) = \tilde{\theta}_{y,b} = E(\theta|y) + \text{cov}(\theta, b|y)\text{var}(b|y)^{-1} \left\{ \tilde{b} - E(b|y) \right\}, \quad (3)$$

with $\tilde{b} = E(b|\theta)$, \bar{F} the $sN \times pN$ matrix $(\bar{A}'\bar{A})^{-1}\bar{A}$, and \bar{A} is the $pN \times sN$ matrix $A \otimes I_N$. Then, $\text{var}(\theta|y, b) = \text{var}(\theta|y) - \text{cov}(\theta, b|y)\text{var}(b|y)^{-1}[\text{cov}(\theta, b|y)]'$.

The predictor $\tilde{\theta}_{y,b}$ in 3 gives the Best Linear Unbiased Predictor $E(\theta|y)$, “embedding” the PPCs through an adjoint component. The last due to knowing that the random vector θ lies in the s -dimensional subspace of the PPCs. In particular, the difference $\tilde{b} - E(b|y)$ gives the multivariate vector of the PPCs “not explained” by the estimation of the linear model $E(\theta|y)$. The matrix $\text{var}(\theta|y, b)$ has rank s , and, consequently, there are $(p - s)$ linear combinations of θ for which their respective variances do not depend on the PPCs.

Proposition 2. *Given the p -dimensional random vector θ , and under the models in (1) and (2), and the Best Predictor $E(\theta|y, b)$ in 3, we get:*

$$\tilde{b}^* = \bar{F} E(\theta|y) + \bar{F} \text{cov}(\theta, b|y)\text{var}(b|y)^{-1} \left\{ \tilde{b} - E(b|y) \right\}, \quad (4)$$

where $\tilde{b}^* = \bar{F} E(\theta|y, b) = \bar{F} \tilde{\theta}_{y,b}$ is the s -dimensional vector of the PPCs “shifted” (SPCs) by the linear predictor $E(\theta|y)$. As a particular case, when $\sigma_\varepsilon^2 \rightarrow 0$, $\text{var}(\varepsilon) \rightarrow 0$, $\text{var}(\gamma) \rightarrow \text{var}(e)$, and $\tilde{b} \rightarrow \bar{b}$. Therefore, $\bar{F} \tilde{\theta}_{y,b} = \tilde{b}^* \rightarrow \bar{b}$, with $\bar{B} = \text{vec}^{-1}(\bar{b}) = ((\text{vec}I_s)' \otimes I_N)(I_s \otimes \bar{b})$ the matrix of the PC scores by sample PCs $\bar{b} = A'\theta$. The matrix $\tilde{B}^* = \text{vec}^{-1}(\tilde{b}^*) = ((\text{vec}I_s)' \otimes I_N)(I_s \otimes \tilde{b}^*)$ can be written as $\text{col}_{1 \leq i \leq n}(\tilde{b}_i^*(t))$, with $\tilde{b}_i^*(t) = \text{col}_{1 \leq t \leq T}(\tilde{b}_{it}^*)$. Thus, the SBCs are the PPCs with the scores predicted by the linear model (1), with subjects as time repeated observations, and $\tilde{b}_s^*(t)$ expressed in terms of spline basis.

The SPCs \tilde{b}^* are then the PPCs “adjusted” by the Linear BP $E(\theta|y)$, and \tilde{b}^* is then the vector obtained by stacking columns of the matrix \tilde{B}^* of the SPC scores. Given $\sigma_\varepsilon^2 = 0$, the vector θ in the model 2 lies in the s -dimensional subspace of

the standard sample PCs \bar{b} . In fact, in this case $\tilde{b} \equiv \bar{b}$, $cov(\theta, b|y) = var(\theta|y)\bar{F}'$, $var(b|y) = \bar{F}var(\theta|y)\bar{F}'$, $E(b|y) = \bar{F}E(\theta|y)$, and then $\bar{F}E(\theta|y, b) = \bar{b}$.

3 Application and Discussion

The data for the application are the Equitable and Sustainable Well-being indicators (in Italian, BES) - annually provided by the Italian Statistical Institute ([3], 2017). In order to highlight the result of the proposed method we use 12 BES indicators relating to the years 2013-2016, collected at NUTS-2 (Nomenclature of Territorial Units for Statistics -2) level. Please, refer to Table 1 for the description and acronyms used for the variables. We use LBE1 as a unique covariate in the LMM model, while the remaining 11 variables are dependent variables. The application uses the REML estimation, a Sas/IML code together with a sequence of Sas-HPMixed procedures. We propose to estimate the model under a uniform correlation structure among the multivariate components of the random effects, and an AR(1) within-subject errors. The uniform correlation between the multivariate components of the random effects is equivalent to compound-symmetry covariance, with a better numerical property in terms of optimization. Further, some studies highlight that using uniform correlation matrices reduces the estimation noise. The slope parameter estimates from the multivariate regression are all significant. The MANOVA multivariate test statistics (Wilks Lambda, Pillai's Trace, Hotelling-Lawley Trace, Roy's Largest Root), based on the characteristic roots are all significant. Figure 1 shows the application of the model described in (3) and (4). The plots represent the spline curves of PPCs and SPCs, for each of the subjects (the 20 Italian administrative Regions). The PPC1 is mainly related with the variables INN1, BS3, AMB9, POL5, S8, while the PPC2 is related with REL4, Q2, L12. The smoothing operated by the linear model part of the SPCs is highlighted especially comparing PPC2 and SPC2, for which the SPC2 shows a less confusing behavior when varying the year. The resulting splines indicate that the prediction stabilizes a lot the scores of the principal components. This is because the SPCs use PPCs, retaining all the main features of the principal components analysis approach, and utilizing at the same time all the capabilities of a linear predictor.

The SPCs in 4 can be viewed from two different perspectives. An "adjusted" linear predictor by the PPCs, by relation (3), as the shifted PCs (SPCs) that modify the probabilistic PCs (PPCs) to accommodate the mixed model regression predicted values. The present work considers then the probabilistic principal components as a "constraint" model, that links together the components of the multivariate random vector in a lower dimensional subspace.

The SPCs are generated considering the multivariate linear predictor 1, by adjusting its standard formulation through the sample parameter vector scores in a convenient subspace. The main advantage of the method proposed is that allows using all the features that a linear model can contribute to improving the principal component data analysis.

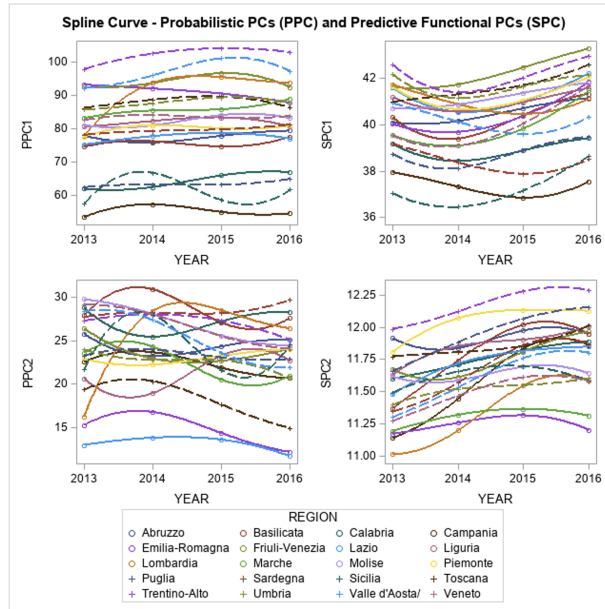


Fig. 1 Joint spline curves of the first two PPCs and SPCs, for the four years (2013-2016) and the twenty Italian administrative Regions. The official Italian Well-Being data are reported and downloadable by the Italian National Statistical Institute web site.

Table 1 Description of the variables used for the application

Variables	Description
S8	Age-standardised mortality rate for dementia and nervous system diseases
IF3	People having completed tertiary education (30-34 years old)
L12	Share of employed persons who feel satisfied with their work
REL4	Social participation
POL5	Trust in other institutions like the police and the fire brigade
SIC1	Homicide rate
BS3	Positive judgment for future perspectives
PATR9	Presence of Historic Parks/Gardens and other Urban Parks recognised of significant public interest
AMB9	Satisfaction for the environment - air, water, noise
INN1	Percentage of R&D expenditure on GDP
Q2	Children who benefited of early childhood services
BE1	Per capita adjusted disposable income
LBE1	Logarithm of Per capita adjusted disposable income

References

1. Demidenko, E.: *Mixed Models: Theory and Applications*. Wiley, New York (2004)
2. Hardle, W. K., Simar, L.: *Principal Components Analysis*. In Hardle, W. K., Simar, L. (eds.) *Applied Multivariate Statistical Analysis*, 319-358. Springer (2015)
3. ISTAT: BES project www.istat.it/en/well-being-and-sustainability
4. Jolliffe Ian T., Cadima J.: *Principal component analysis: a review and recent developments* *Phil. Trans. R. Soc. A.* 374: 20150202.20150202 (2016)
5. Longford N.T.: *Random Coefficient Models*. In: Lovric M. (eds.) *International Encyclopedia of Statistical Science*. Springer, Heidelberg (2011)
6. Timm, N. H.: *Applied Multivariate Analysis*. Springer, New York (2002)
7. Tipping, M.E., Bishop C.M.: *Probabilistic principal component analysis*. *J. R. Stat. Soc., Ser. B (Stat. Methodol.)* 61(3), 611–622 (1999)

This book collects the papers presented at the 11th Scientific Meeting of the SIS Group “Statistics for the Evaluation and Quality in Services” *Statistical Methods for Evaluation and Quality: Techniques, Technologies and Trends (T^B)*, which took place at the University of Chieti-Pescara, on 30th August-1st September, 2023.

The papers, which had been selected through a refereeing process, contain topics on statistical approaches and methodologies for the evaluation of public services in different contexts, and cover the areas of digital transition, e-commerce and digital marketing, enterprises, environment and territory, healthcare and wellness, finance, bank and FinTech, justice system, labour market, official statistics, public administration, food and wine, school, education and training, social, sports, sustainability, tourism, transport, university and research, well-being and welfare.

ISBN: 979-12-803-3369-8

DOI: 10.60984/978-88-94593-36-5-IES2023

ISBN 9791280333698



9 791280 333698 >