


Smart Statistics for Smart Applications

Book of Short Papers SIS2019



Editors: Giuseppe Arbia, Stefano Peluso,
Alessia Pini and Giulia Rivellini

Copyright © 2019

PUBLISHED BY PEARSON

WWW.PEARSON.COM

Giugno 2019 ISBN 9788891915108

Preface

Section 1. Plenary Sessions and Round Table

Preface	3
Shallow Learning for Data Science	7
<i>Antonio Canale</i>	
Smart Statistics: concept, technology and service	17
<i>David John Hand, Maurizio Vichi</i>	
Tavola rotonda “Smart ageing: lunga vita attiva, salute e nuove tecnologie”	19

Section 2. Invited Papers

Demography in the Digital Era: New Data Sources for Population Research	23
Demografia nell’era digitale: nuovi fonti di dati per gli studi di popolazione	23
<i>Diego Albrez-Gutierrez, Samin Aref, Sofia Gil-Clavel, André Grow, Daniela V. Negraia, Emilio Zagheni</i>	
Stationarity of a general class of observation driven models for discrete valued processes	31
Stazionarietà di una classe generale di modelli observation-driven per processi a valori discreti	
<i>Mirko Armillotta, Alessandra Luati and Monia Lupporelli</i>	
An extension of the censored gaussian lasso estimator	39
Un’estensione dello stimatore cglasso	
<i>Luigi Augugliaro and Gianluca Sottile and Veronica Vinciotti</i>	
A formal approach to data swapping and disclosure limitation techniques	47
Un approccio formale per tecniche di trasformazione dei dati in problemi di privacy	
<i>F. Ayed, M. Battiston and F. Camerlenghi</i>	
A new ordinary kriging predictor for histogram data in L2-Wasserstein space	55
Un nuovo predittore kriging per istogrammi nello spazio L2-Wasserstein	
<i>Antonio Balzanella and Antonio Irpino and Rosanna Verde</i>	
Keywords dynamics in online social networks: a case-study from Twitter	63
La dinamica delle parole chiave nelle reti sociali online: un esempio tratto da Twitter	
<i>Carolina Becatti, Irene Crimaldi and Fabio Saracco</i>	
Statistical Matching of HBS and ADL to analyse living conditions, poverty and happiness	71
Statistical Matching di HBS e ADL per l’analisi di condizioni di vita, povertà e felicità	
<i>Cristina Bernini, Silvia Emili, Maria Rosaria Ferrante</i>	
Statistical sources for cybersecurity and measurement issues	79
Fonti statistiche per la sicurezza cibernetica e problemi di misurazione	
<i>Claudia Biancotti, Riccardo Cristadoro, Raffaele Tartaglia Polcini</i>	
Use of GPS-enabled devices data to analyse commuting flows between Tuscan municipalities	89
Un’analisi dei flussi di pendolarismo sistematici tra i comuni toscani tramite l’utilizzo di dati GPS	
<i>Chiara Bocci, Leonardo Piccini and Emilia Rocco</i>	
Statistical calibration of the digital twin of a connected health object	97
Inversione statistica dei parametri di ingresso per il gemello digitale di un oggetto sanitario collegato	
<i>Nicolas Bousquet and Walid Dabachine</i>	
Time Series Forecasting: Is there a role for neural networks?	103
Le Reti Neuronali nella Previsione di Serie Storiche	
<i>Giuseppe Bruno, Sabina Marchetti, Juri Marcucci, Diana Nicoletti</i>	

Modelling weighted signed networks.....	111
Modellazione di reti segnate pesate	
<i>Alberto Caimo and Isabella Gollini</i>	
Issues on Bayesian nonparametric measures of disclosure risk	119
Questioni su misure Bayesiane nonparametriche di rischio di "disclosure"	
<i>Federico Camerlenghi, Cinzia Carota and Stefano Favaro</i>	
Hierarchies of nonparametric priors.....	125
Gerarchie di distribuzioni iniziali nonparametriche	
<i>Federico Camerlenghi, Stefano Favaro and Lorenzo Masoero</i>	
Issues with Nonparametric Disclosure Risk Assessment.....	133
Questioni sull'Analisi Nonparametrica del Rischio di "Disclosure"	
<i>Federico Camerlenghi, Stefano Favaro, Zacharie Naulet and Francesca Panero</i>	
Technologies and data science for a better health both at individual and population level. ..	141
Two practical research cases.	141
Tecnologie e data science per una salute migliore sia a livello individuale che di popolazione.	
<i>Stefano Campostrini and Lucia Zanotto</i>	
Temporal sentiment analysis with distributed lag models	149
Analisi temporale del "sentiment" con modelli a lag distribuiti	
<i>Carrannante M., Mattered R., Misuraca M., Scepi G., Spano M.</i>	
A statistical investigation on the relationships among financial disclosure, sociodemographic variables, financial literacy and retail investors' risk assessment ability	157
Indagine empirica sulle relazioni tra prospetti per la diffusione di informazioni finanziarie, variabili sociodemografiche, educazione finanziaria e abilità di valutazione del rischio	
<i>Rosella Castellano, Marco Mancinelli and Pasquale Sarnacchiaro</i>	
Bayesian Model Comparison based on Wasserstein Distances.....	167
Confronto di Modelli Bayesiani tramite Distanze di Wasserstein	
<i>Marta Catalano, Antonio Lijoi and Igor Prünster</i>	
Hierarchical Clustering and Dimensionality Reduction for Big Data	173
Clustering e Riduzione Dimensionale Gerarchici per Dati di Grandi Dimensioni	
<i>Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria</i>	
ICOs success drivers: a textual and statistical analysis.....	181
Fattori di successo nelle ICOs: un'analisi testuale e statistica	
<i>Paola Cerchiello and Anca Mirela Toma</i>	
Small area estimators with linked data.....	189
Stimatori per piccole aree nel caso di dati ottenuti attraverso il record linkage	
<i>Chambers Raymond and Fabrizi Enrico and Salvati Nicola</i>	
Optimal Portfolio Selection via network theory in banking and insurance sector.....	197
<i>Gian Paolo Clemente, Rosanna Grassi and Asmerilda Hitaj</i>	
Matching error(s) and quality of statistical matching in complex surveys.....	205
Errori di matching e qualità del matching statistico in indagini complesse	
<i>Pier Luigi Conti and Daniela Marella</i>	
Hotel search engine architecture based on online reviews' content.....	213
Un motore di ricerca per gli hotel basato sulle recensioni online	
<i>Claudio Conversano, Maurizio Romano and Francesco Mola</i>	
Economic Crisis and Earnings Management: a Statistical Analysis	219
Crisi Economica e Gestione degli Utili: un'Analisi Statistica	
<i>C. Cusatelli, A.M. D'Uggento, M. Giacalone, F. Grimaldi</i>	
A Comparison of Nonparametric Bivariate Survival Functions.....	227
Confronto tra stimatori non-parametrici della funzione di sopravvivenza bivariata	
<i>Hongsheng Dai and Marialuisa Restaino</i>	
Predictive Algorithms in Criminal Justice.....	237
Algoritmi predittivi e giustizia penale	
<i>Francesco D'Alessandro</i>	

A proposal for an integrated approach between sentiment analysis and social network analysis.....	247
Una proposta per un approccio integrato tra analisi del sentimento e analisi delle reti sociali	
<i>Domenico De Stefano and Francesco Santelli</i>	
A meta-tissue non-parametric factor analysis model for gene co-expression	255
Meta-analisi fattoriale non parametrica per lo studio di espressioni genetiche in diversi tessuti	
<i>Roberta De Vito and Barbara Engelhardt</i>	
Bayesian estimate of population count with false captures: a latent class approach	261
Stima Bayesiana della popolazione con false catture: un approccio basato sulle classi latenti	
<i>Davide Di Cecco, Marco Di Zio and Brunero Liseo</i>	
Spherical regression with local rotations and implementation in R	269
Regressione sferica con rotazioni locali ed implementazione in R	
<i>Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor</i>	
A clustering method for network data to analyse association football playing styles	277
Un metodo di raggruppamento per dati di rete finalizzato all'analisi degli schemi di gioco nel calcio	
<i>Jacopo Diquigiovanni</i>	
Big data in longitudinal observational studies: how to deal with non-probability samples and technological changes.....	285
I Big data negli studi longitudinali: come trattare campioni non probabilistici e cambi di tecnologia	
<i>Clelia Di Serio, Luca Del Core, Eugenio Montini and Andrea Calabria</i>	
Smart Data For Smart Health.....	293
Smart Data Per Smart Health	
<i>Clelia Di Serio, Ernst C. Wit, Elena Bottinelli and Roberto Buccione</i>	
Detecting and classifying moments in basketball matches using sensor tracked data.....	297
Una procedura per identificare e classificare momenti di gioco in pallacanestro con l'uso di dati sensori.	
<i>Tullio Facchinetti and Rodolfo Metulini and Paola Zuccolotto</i>	
Ordered response models for cyber risk	305
Modelli a risposta ordinale per la valutazione del cyber risk	
<i>Silvia Facchinetti and Claudia Tarantola</i>	
Functional data analysis-based sensitivity analysis of integrated assessment Models for climate change modelling	313
Analisi di sensibilità basata sull'analisi di dati funzionali per modelli di valutazione integrata dei cambiamenti climatici	
<i>Matteo Fontana, Massimo Tavoni and Simone Vantini</i>	
Coupled Gaussian Processes for Functional Data Analysis.....	319
Processi gaussiani per l'analisi dei dati funzionali	
<i>L. Fontanella, S. Fontanella, R. Ignaccolo, L. Ippoliti, P. Valentini</i>	
Two-fold data streams dimensionality reduction approach via FDA	323
Un approccio a due fasi per la riduzione di dimensionalità di data streams via FDA	
<i>F. Fortuna, T. Di Battista and S.A. Gattone</i>	
Statistical analysis of Sylt's coastal profiles using a spatiotemporal functional model.....	331
<i>Rik Gijssman, Philipp Otto, Torsten Schlurmann, Jan Visscher</i>	
Bootstrap prediction intervals for weighted TAR predictors	339
Intervalli di previsione bootstrap per previsori ponderati per modelli TAR	
<i>Francesco Giordano and Marcella Niglio</i>	
A rank graduation index to prioritise cyber risks	347
Un indice di graduazione per assegnare livelli di priorità ai rischi informatici	
<i>Paolo Giudici and Emanuela Raffinetti</i>	
Vector Error Correction models to measure connectedness of bitcoin exchange markets	355
Modelli di Vector Error Correction per misurare la connessione delle piattaforme di scambio di bitcoin	
<i>Paolo Giudici and Paolo Pagnottoni</i>	
Estimation of lineup efficiency effects in Basketball using play-by-play data.....	363
L'uso dei dati del play-by-play per la stima degli effetti di quintetto nella pallacanestro	
<i>Luca Grassetti, Ruggero Bellio, Giovanni Fonseca and Paolo Vidoni</i>	
Trajectory clustering using adaptive squared distances	371
Clustering di traiettorie attraverso distanze adattative quadratiche	
<i>Antonio Irpino</i>	

Bayesian Analysis of Privacy Attacks on GPS Trajectories	379
<i>Analisi Bayesiana degli Attacchi alla Privacy su Traiettorie GPS</i>	
<i>Sirio Legramanti</i>	
Data Analytics in the Insurance Industry: Market trends and lessons from a use case customer predictive modelling	387
<i>Data Analytics nel settore assicurativo: principali trend e considerazioni da un caso d'uso applicato alla predizione del comportamento degli assicurati</i>	
<i>Cristian Losito and Francesco Pantisano</i>	
BasketballAnalyzeR: the R package for basketball analytics	395
<i>BasketballAnalyzeR: il pacchetto R per l'analisi dei dati nella pallacanestro</i>	
<i>Marica Manisera, Marco Sandri and Paola Zuccolotto</i>	
Data Integration by Graphical Models.....	403
<i>Utilizzo dei modelli grafici per l'integrazione dei dati</i>	
<i>Daniela Marella and Paola Vicard and Vincenzina Vitale</i>	
A two-part finite mixture quantile regression model for semi-continuous longitudinal data	409
<i>Maruotti Antonello, Merlo Luca and Petrella Lea</i>	
Multivariate change-point analysis for climate time series	415
<i>Analisi di change-point multivariate per serie storiche climatiche</i>	
<i>Gianluca Mastrantonio, Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio and Carlo Biasi</i>	
A divide-et-impera approach for the spatial prediction of object data over complex regions.....	423
<i>Un approccio divide-et-impera per la previsione spaziale di dati oggetto su regioni complesse</i>	
<i>Alessandra Menafoglio e Piercesare Secchi</i>	
A strategy for the matching of mobile phone signals with census data.....	427
<i>Una strategia per l'abbinamento di segnali di telefonia mobile con dati censuari</i>	
<i>Rodolfo Metulini and Maurizio Carpita</i>	
Risk-based analyses for non-proportional reinsurance pricing	435
<i>Analisi Risk-based per il pricing nella riassicurazione di trattati non proporzionali</i>	
<i>Fabio Moraldi and Nino Savelli</i>	
A Simplified Efficient and Direct Unequal Probability Resampling	441
<i>Un semplice Ricampionamento, efficiente e diretto per campioni a probabilità variabili</i>	
<i>Federica Nicolussi, Fulvia Mecatti and Pier Luigi Conti</i>	
Labour Law: Machine vs. Employer Powers Diritto del lavoro: Macchina vs. Poteri datoriali	449
<i>Antonella Occhino – Michele Faioli</i>	
Domain knowledge based priors for clustering.....	455
<i>Distribuzioni a priori per l'analisi di raggruppamento basate sulla conoscenza di settore</i>	
<i>Sally Paganin</i>	
Clustering of Behavioral Spatial Trajectories in Neuropsychological Assessment	463
<i>Analisi dei gruppi di traiettorie spaziali nella valutazione neuropsicologica</i>	
<i>Francesco Palumbo, Antonio Cerrato, Michela Ponticorvo, Onofrio Gigliotta, Paolo Bartolomeo, Orazio Miglino</i>	
What is wrong in the debate about smart contracts.....	471
<i>Smart contract e diritto: riflessioni critiche su un dualismo fuorviante</i>	
<i>Roberto Pardolesi and Antonio Davola</i>	
Financial Transaction Data for the Nowcasting in Official Statistics.....	485
<i>Transazioni elettroniche di pagamento per le previsioni a breve nella Statistica ufficiale</i>	
<i>Righi A., Ardizzi G., Gambini A., Iannaccone R., Moauro F., Renzi N. and Zurlo D.</i>	
On the examination of a criticality measure for a complex system in a forecasting perspective	493
<i>Esame di una misura di criticità per un sistema complesso in una prospettiva previsiva</i>	
<i>Renata Rotondi and Elisa Varini</i>	
Knowledge discovery for dynamic textual data: temporal patterns of topics and word clusters in corpora of scientific literature	501
<i>Estrazione della conoscenza da dati testuali dinamici: evoluzione temporale di argomenti e gruppi di parole in corpora di letteratura scientifica</i>	
<i>Stefano Sbalchiero, Matilde Trevisani and Arjuna Tuzzi</i>	

Classifying the Willingness to Act in Social Media Data: Supervised Machine Learning for U.N. 2030 Agenda	509
Classificare la volontà di agire nei dati dei Social Media: Supervised Machine Learning per l'Agenda 2030 delle Nazioni Unite	
<i>Andrea Sciandra, Alessio Surian and Livio Finos</i>	
Classification of spatio-temporal point pattern in the presence of clutter using K-th nearest neighbour distances.....	517
Classificazione dei processi puntuali spazio-temporali basata sulla distanza dal K-mo vicino più vicino	
<i>Siino Marianna, Francisco J. Rodriguez-Cortés, Jorge Mateu, Giada Adelfio</i>	
Modelling properties of high-dimensional molecular systems	525
La modellazione di sistemi molecolari ad alta dimensionalità	
<i>Debora Slanzi, Valentina Mamelì and Irene Poli</i>	
Non-crossing parametric quantile functions: an application to extreme temperatures	533
Il problema del crossing con funzioni quantiliche parametriche: un'applicazione alle temperature estreme	
<i>Gianluca Sottile and Paolo Frumento</i>	
A new tuning parameter selector in lasso regression.....	541
Un nuovo criterio di selezione per il parametro di penalizzazione nella regressione lasso	
<i>Gianluca Sottile and Vito MR Muggeo</i>	
Similarity patterns, topological information and credit scoring models	549
Strutture di similarità, informazioni topologiche e modelli di credit scoring	
<i>Alessandro Spelta, Branka Hadji-Misheva and Paolo Giudici</i>	
Between hawks and doves: measuring central bank communication	557
Fra falchi e colombe: valutazione delle comunicazioni di Banca Centrale	
<i>Ellen Tobback, Stefano Nardelli, David Martens</i>	
New methods and data sources for the population census	561
Nuovi metodi e fonti per il censimento della popolazione	
<i>Paolo Valente</i>	
FinTech and the Search for "Smart" Regulation	569
Fintech e la ricerca di una regolamentazione "smart"	
<i>Silvia Vanon</i>	
An anisotropic model for global climate data	577
Un modello anisotropico per i dati climatici globali	
<i>Nil Venet and Alessandro Fassò</i>	
Analysis of the financial performance in Italian football championship clubs via GEE and diagnostic measures.....	585
Analisi delle performance finanziaria delle squadre di calcio di serie A via GEE e misure di diagnostica	
<i>Maria Kelly Venezuela, Anna Crisci, Luigi D'Ambra, D'Ambra Antonello</i>	
A statistical space-time functional model for air quality analysis and mapping.....	593
Un modello statistico spazio-tempo funzionale per l'analisi e la mappatura della qualità dell'aria	
<i>Yaqiong Wang, Alessandro Fassò and Francesco Finazzi</i>	
Tempering and computational efficiency of Bayesian variable selection.....	599
Tempering e l'efficienza computazionale della selezione bayesiana delle variabili	
<i>Giacomo Zanella and Gareth O. Roberts</i>	
Dimensions and links for Hate Speech in the social media	607
Dimensioni e legami per i discorsi di odio nei social media	
<i>Emma Zavarrone, Guido Ferilli</i>	

Section 3. Contributed Papers

Density-based Algorithm and Network Analysis for GPS Data.....	617
Algoritmi di Cluster e Reti per lo studio di dati GPS	
<i>Antonino Abbruzzo, Mauro Ferrante, Stefano De Cantis</i>	
Local inference on functional data based on the control of the family-wise error rate	623
Inferenza locale per dati funzionali basata sul controllo del family-wise error rate	
<i>Konrad Abramowicz, Alessia Pini, Lina Schelin, Sara Sjöstedt de Luna, Aymeric Stamm, and Simone Vantini</i>	

Application and validation of dynamic Poisson models to measure credit contagion	629
<i>Applicazione e validazione di modelli di Poisson dinamici per misurare il contagio nel credito</i>	
<i>Arianna Agosto and Emanuela Raffinetti</i>	
Monitoring SDGs at territorial level: the case of Lombardy.....	637
<i>Il monitoraggio degli SDGs a livello territoriale: il caso della Lombardia</i>	
<i>Leonardo Alaimo, Livia Celardo, Filomena Maggino, Adolfo Morrone, Federico Olivieri</i>	
The Experts Method for the prediction of periodic multivariate time series of high dimension.....	643
<i>Il Metodo degli Esperti per la previsione di serie temporali multivariate e periodiche, di dimensione elevata</i>	
<i>Giacomo Aletti, Marco Bellan and Alessandra Micheletti</i>	
Regression with time-dependent PDE regularization for the analysis of spatio-temporal data	649
<i>Regressione con regolarizzazione di PDE tempo dipendenti per modellizzare dati spazio-temporali</i>	
<i>Eleonora Arnone, Laura Azzimonti, Fabio Nobile, Laura M. Sangalli</i>	
A network analysis of museum preferences: the Firenzecard experience.....	653
<i>Un'analisi di rete delle preferenze museali: l'esperienza della Firenzecard</i>	
<i>Silvia Bacci, Bruno Bertaccini, Roberto Dinelli, Antonio Giusti, and Alessandra Petrucci</i>	
A statistical learning approach to group response categories in questionnaires.....	659
<i>Un approccio basato sull'apprendimento statistico per raggruppare le categorie di risposta nei questionari</i>	
<i>Michela Battauz</i>	
Tree-based Functional Data Analysis for Classification and Regression.....	665
<i>Alberi di Classificazione e Regressione per dati Funzionali</i>	
<i>Edoardo Belli, Enrico Ragaini, Simone Vantini</i>	
PDE-regularized regression for anisotropic	669
<i>spatial fields Regressione con regolarizzazione differenziale per campi spaziali anisotropi</i>	
<i>Mara S. Bernardi, Michelle Carey, James O. Ramsay and Laura M. Sangalli</i>	
A Bayesian model for network flow data: an application to BikeMi trips	673
<i>Giulia Bissoli, Celeste Principi, Gian Matteo Rinaldi, Mario Beraha and Alessandra Guglielmi</i>	
Statistical classics in the big data era. When (astro-physical) models are nonregular.....	679
<i>Statistica classica nell'era dei big data. Verosimiglianza e modelli non regolari</i>	
<i>Alessandra R. Brazzale and Valentina Mameli</i>	
Bayesian Variable Selection for High Dimensional Logistic Regression	685
<i>Selezione bayesiana delle variabili nel modello di regressione logistica ad alta dimensionalita</i>	
<i>Claudio Busatto, Andrea Sottosanti and Mauro Bernardi</i>	
Bayesian modeling for large spatio-temporal data: an application to mobile networks	691
<i>Modelli bayesiani per grandi dataset spazio-temporali: un'applicazione a dati di telefonia mobile</i>	
<i>Annalisa Cadonna, Andrea Cremaschi, Alessandra Guglielmi</i>	
A Mathematical Framework for Population of Networks: Comparing Public Transport of Different Cities.	697
<i>Un approccio matematico all'analisi di una popolazione di networks: come confrontare il sistema di trasporto pubblico di diverse città.</i>	
<i>Anna Calissano, Aasa Feragen, Simone Vantini</i>	
How Important Discrimination is for the Job Satisfaction of Immigrants in Italy: A Counterfactual Approach.....	703
<i>Quanto influisce la discriminazione sulla soddisfazione lavorativa degli immigrati in Italia: un approccio controfattuale</i>	
<i>Maria Gabriella Campolo, Antonino Di Pino and Michele Limosani</i>	
Unfolding the SEcrets of LongEvity: Current Trends and future prospects (SELECT)	709
<i>A path through morbidity, disability and mortality in Italy and Europe</i>	
<i>Stefano Campostrini, Daniele Durante, Fabrizio Faggiano and Stefano Mazzucco</i>	
Galaxy color distribution estimation via dependent nonparametric mixtures	713
<i>Stima della distribuzione del colore delle galassie via misture nonparametriche dipendenti</i>	
<i>Antonio Canale, Riccardo Corradin and Bernardo Nipoti</i>	
A case for order optimal matching: a salary gap study.....	719
<i>Un algoritmo di matching ottimale ordinato per un studio sulle differenze salariali</i>	
<i>Massimo Cannas</i>	

A Prediction Method for Ordinal Consistent Partial Least Squares	725
Un Metodo di Previsione per l'Algoritmo Ordinal Consistent Partial Least Squares	
<i>Gabriele Cantaluppi and Florian Schubert</i>	
Functional control charts for monitoring ship operating conditions and CO2 emissions based on scalar-on-function linear model	731
Carte di controllo funzionali per il monitoraggio delle condizioni operative e delle emissioni di CO2 di navi da carico e passeggeri mediante modello di regressione funzionale con risposta scalare	
<i>Christian Capezza, Antonio Lepore, Alessandra Menafoglio, Biagio Palumbo, and Simone Vantini</i>	
Predicting and improving smart mobility: a robust model-based approach to the BikeMi BSS	737
Prevedere e migliorare la mobilità smart: un approccio robusto di classificazione applicato a BikeMi	
<i>Andrea Cappozzo, Francesca Greselin and Giancarlo Manzi</i>	
Public support for an EU-wide social benefit scheme: evidence from Round 8 of the European Social Survey (ESS)	743
Sostegno pubblico a un sistema di prestazioni sociali a livello dell'Unione Europea: i risultati del Round 8 della European Social Survey (ESS)	
<i>Paolo Emilio Cardone</i>	
Revenue management strategies and Booking.com ghost rates: a statistical analysis	751
Strategie di revenue management e Booking.com ghost rates: un'analisi statistica	
<i>Cinzia Carota, Consuelo R. Nava, Marco Alderighi</i>	
Analysing international migration flows: a Bayesian network approach	757
Analisi dei flussi migratori internazionali attraverso l'impiego di modelli grafici	
<i>Federico Castelletti and Emanuela Furfaro</i>	
A sparse estimator for the function-on-function linear regression model	763
Uno stimatore sparso per il modello di regressione lineare con regressore e risposta funzionali	
<i>Fabio Centofanti, Matteo Fontana, Antonio Lepore, and Simone Vantini</i>	
Robustness and fuzzy multidimensional poverty indicators: a simulation study.....	769
Robustezza ed indicatori fuzzy multidimensionali della povertà: uno studio di simulazione	
<i>Michele Costa</i>	
Text Based Pricing Modelling: an Application to the Fashion Industry	775
Modellazione dei prezzi basata su dati testuali: un'applicazione all'industria fashion	
<i>Federico Crescenzi, Marzia Freo and Alessandra Luati</i>	
Model based clustering in group life insurance via Bayesian nonparametric mixtures	781
Raggruppamento basato sul modello nel settore assicurativo: un approccio bayesiano nonparametrico	
<i>Laura D'Angelo</i>	
Smart Tools for Academic Submission Decisions: Waiting Times Modeling	787
Strumenti "Smart" per sottoporre i manoscritti accademici: modelli per i tempi di attesa	
<i>Francesca De Battisti - Giancarlo Manzi</i>	
On the Use of Control Variables in PLS-SEM	793
Sull'Uso delle Variabili di Controllo nei PLS-SEM	
<i>Francesca De Battisti and Elena Siletti</i>	
Partial dependence with copula and financial applications	799
Dipendenza parziale con funzioni copula e applicazioni finanziarie	
<i>Giovanni De Luca, Marta Nai Ruscone and Giorgia Riveccio</i>	
Exploring the relationship between fertility and well-being: What is smart?.....	805
Esplorando la relazione tra fecondità e benessere: cosa c'è di smart?	
<i>Alessandra De Rose, Filomena Racioppi, Maria Rita Sebastiani</i>	
Web-Based Data Collection and Quality Issues in Co-Authorship Network Analysis	811
Qualità dei dati bibliografici raccolti via web per l'analisi di reti di collaborazione scientifica	
<i>Domenico De Stefano, Vittorio Fuccella, Susanna Zaccarin</i>	
A new regression model for bounded multivariate responses.....	817
Un nuovo modello di regressione per risposte multivariate limitate	
<i>Agnese Maria Di Brisco, Roberto Ascari, Sonia Migliorati and Andrea Ongaro</i>	
Turning big data into smart data: two examples based on the analysis of the Mappa dei Rischi dei Comuni Italiani.....	823
Trasformare i big data in smart data: due esempi di analisi della Mappa dei Rischi dei Comuni Italiani	
<i>Oleksandr Didkovskiy, Alessandra Menafoglio, Piercesare Secchi, Giovanni Azzone</i>	

Hidden Markov Model estimation via Particle Gibbs	829
Stima di Hidden Markov Model tramite Particle Gibbs	
<i>Pierfrancesco Alaimo Di Loro, Enrico Ciminello and Luca Tardella</i>	
A note on marginal effects in logistic regression with independent covariates	837
Una nota sugli effetti marginali nella regressione logistica con covariate indipendenti	
<i>Marco Doretti</i>	
DNA mixtures: a case study involving a Romani reference population.....	843
Misture di DNA: un caso di studio riguardante una popolazione di riferimento dei Rom	
<i>Francesco Dotto, Julia Mortera and Vincenzo Pascali</i>	
Pivotal seeding for K-means based on clustering ensembles	849
Inizializzazione pivotale dell'algoritmo delle K-medie tramite raggruppamento con metodi di insieme	
<i>Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli</i>	
Optimal scoring of partially ordered data, with an application to the ranking of smart cities	855
Scoring ottimale di dati parzialmente ordinati, con un'applicazione al ranking delle smart city	
<i>Marco Fattore, Alberto Arcagni, Filomena Maggino</i>	
Bounded Domain Density Estimation	861
Stima della densità non-parametrica su domini bidimensionali limitati	
<i>Federico Ferraccioli, Laura M. Sangalli and Livio Finos</i>	
Polarization and long-run mobility: yearly wages comparison in three southern European countries.....	867
Polarizzazione e mobilità sul lungo periodo: un confronto fra salari annuali in tre Paesi sud-Europei	
<i>Ferretti C., Crosato L., Cipollini F., Ganugi P.</i>	
Design of Experiments, aberration and Market Basket Analysis.....	873
Pianificazione degli esperimenti, aberrazione e Market Basket Analysis	
<i>Roberto Fontana and Fabio Rapall</i>	
Generalized Procrustes Analysis for Multilingual Studies	879
Analisi Procrustiana Generalizzata per studi Multilingue	
<i>Alessia Forciniti, Michelangelo Misuraca, Germana Scepi, Maria Spano</i>	
Prior specification in flexible models.....	885
Specificazione delle prior in modelli flessibili	
<i>Maria Franco-Villoria, Massimo Ventrucci and Haavard Rue</i>	
Modeling Cyclists' Itinerary Choices: Evidence from a Docking Station-Based Bike-Sharing System.....	889
Un modello per gli itinerari dei ciclisti: risultati da un bike-sharing a stazioni fisse	
<i>S. T. Gaito - G. Manzi - G. Saibene - S. Salini - M. Zignani</i>	
A PARAFAC-ALS variant for fitting large data sets	895
Una variante del PARAFAC-ALS per approssimare data set di grandi dimensioni	
<i>Michele Gallo, Violetta Simonacci and Massimo Guarino</i>	
A Convex Mixture Model for Binomial Regression	901
Un modello mistura convessa per la Regressione Binomiale	
<i>Luisa Galtarossa and Antonio Canale</i>	
Blockchain as a universal tool for business improvement	907
Blockchain come strumento universale per il miglioramento del business	
<i>Massimiliano Giacalone, Diego Carmine Sinito, Emilio Massa, Federica Oddo, Enrico Medda, Vito Santarcangelo</i>	
Seasonality in tourist flows: a decomposition of the change in seasonal concentration.....	913
La stagionalità nei flussi turistici: una scomposizione della variazione nella concentrazione stagionale	
<i>Luigi Grossi and Mauro Mussini</i>	
Are Real World Data the smart way of doing Health Analytics?.....	919
Real World Data: la base di una nuova ricerca clinica?	
<i>Francesca Ieva</i>	
Internet use and leisure activities: are all young people equal?.....	925
Internet e tempo libero: i giovani sono uguali tra loro?	
<i>Giuseppe Lamberti, Jordi Lopez Sintas and Pilar Lopez Belbeze</i>	
On a Family of Transformed Stochastic Orders	931
Su una famiglia di ordinamenti stocastici trasformati	
<i>Tommaso Lando and Lucio Bertoli-Barsotti</i>	

Bayesian stochastic search for Ising chain graph models.....	935
<i>Ricerca stocastica Bayesiana per modelli grafici a catena Ising</i>	
<i>Andrea Lazerini · Monia Lupporelli · Francesco C. Stingo</i>	
On the statistical design of parameters for variables sampling plans based on process capability index Cpk	941
<i>Progettazione statistica dei parametri per il piano di campionamento per variabili basato sull'indice di capacità di processo Cpk</i>	
<i>Antonio Lepore, Biagio Palumbo and Philippe Castagliola</i>	
Nowcasting foreign tourist arrivals using Google Trends: an application to the city of Florence, Italy.....	947
<i>Nowcasting degli arrivi turistici stranieri usando Google Trends: un'applicazione nella città di Firenze, Italia</i>	
<i>Alessandro Magrini</i>	
Inclusive growth in European countries: a cointegration analysis	953
<i>La crescita inclusiva nei paesi europei: un'analisi di cointegrazione</i>	
<i>Paolo Mariani, Andrea Marletta, Alessandra Michelangeli</i>	
ESCO- the European Labour Language: a conceptual and operational asset in support of labour governance in complex environments	959
<i>ESCO il linguaggio europeo del lavoro: uno strumento concettuale ed operativo per le politiche del lavoro in contesti complessi</i>	
<i>Cristilla Martelli, Laura Grassini, Adham Kahlawi, Maria Flora Salvatori, Lucia Buzzigoli</i>	
Hidden Markov Models for High Dimensional Data	965
<i>Hidden Markov Models per dati ad alta dimensionalità</i>	
<i>Martino, A., Guatter, G., Paganoni, A.M.</i>	
Classification of Italian classes via bivariate semi parametric multilevel models	971
<i>Classificazione delle classi italiane per mezzo di modelli bivariati a effetti misti semi parametrici</i>	
<i>Chiara Masci, Francesca Ieva, Tommaso Agasisti and Anna Maria Paganoni</i>	
Data Mining Application to Healthcare Fraud Detection: Two-Step Unsupervised Clustering Method for Outlier Detection with Administrative Databases.....	977
<i>Data Mining Applicato al Riconoscimento Frodi in Sanità: Algoritmo a Due Step per l'Identificazione di Outliers con Database Amministrativi</i>	
<i>Massi Michela C., Ieva Francesca, Lettieri Emanuele</i>	
Multivariate analysis and biodiversity partitioning of a demersal fish community: an application to Lazio coast	985
<i>Analisi multivariata e partizione della biodiversità di una comunità di specie demersali: un'applicazione alla costa laziale</i>	
<i>M. Mingione, G. Jona Lasinio, S. Martino, F. Colloca</i>	
Latent Markov models with discrete separate cluster random effects on initial and transition probabilities.....	991
<i>Modelli Latent Markov ad effetti casuali discreti e separati per le probabilità iniziali e di transizione</i>	
<i>Giorgio E. Montanari and Marco Doretti</i>	
Unsuitability of likelihood-based asymptotic confidence intervals for Response-Adaptive designs in normal homoscedastic trials	997
<i>Inadeguatezza degli intervalli di confidenza asintotici basati sulla verosimiglianza per disegni Response-Adaptive in caso di risposte normali omoschedastiche</i>	
<i>Marco Novelli and Maroussa Zagoraiou</i>	
Local Hypothesis Testing for Functional Data: Extending False Discovery Rate to the Functional Framework.....	1003
<i>Verifica locale delle ipotesi nell'ambito dei dati funzionali: estensione della nozione di False Discovery Rate al contesto funzionale</i>	
<i>Niels Asken Lundtorp Olsen, Alessia Pini, and Simone Vantini</i>	
Educational mismatch and attitudes towards migration in Europe.....	1009
<i>Disallineamento fra formazione e lavoro e atteggiamenti verso le migrazioni in Europa</i>	
<i>Marco Guido Palladino and Emiliano Sironi</i>	
Soft thresholding Bayesian variable selection for compositional data analysis.....	1015
<i>Selezione di Variabili Bayesiana con funzioni di soglia per l'analisi di dati di composizione</i>	
<i>Matteo Pedone, Francesco C. Stingo</i>	
Sentiment-driven investment strategies: a practical example of AI-powered engines in a corporate setting	1021
<i>Strategie d'investimento guidate dal sentiment: un esempio pratico di Intelligenza Artificiale in contesto aziendale</i>	
<i>Mattia Pedrini, Sebastian Donoso, Enrico Deusebio, Nicola Donelli, Gabriele Arici, Andrea Cosentini, Paola Mosconi, Diego Ostinelli and Claudio Cocchis</i>	

Betting on football: a model to predict match outcomes	1027
Scommettere sul calcio: un nuovo modello per prevedere l'esito delle partite	
<i>Marco Petretta, Lorenzo Schiavon and Jacopo Diquigiovanni</i>	
Estimation of dynamic quantile models via the MM algorithm	1033
Stima di modelli Quantilici Dinamici con algoritmo MM	
<i>Fabrizio Poggioni, Mauro Bernardi, Lea Petrella</i>	
The decomposition by subpopulations of the Pietra index: an application to the professional football teams in Italy	1039
La scomposizione per sottopopolazioni dell'indice di Pietra: un'applicazione alle squadre professionistiche di calcio in Italia	
<i>Francesco Porro and Mariangela Zenga</i>	
An Object Oriented Data Analysis of Tweets: the Case of Queen Elizabeth Olympic Park.	1045
Object Oriented Data Analysis di Tweet: il caso del Queen Elizabeth Olympic Park	
<i>Paola Riva, Paola Sturla, Anna Calissano and Simone Vantini</i>	
Bias reduced estimation of a fixed effects model for Expected Goals in association football	1051
Stima non distorta di un modello Expected Goal con effetti fissi nel calcio	
<i>Lorenzo Schiavon and Nicola Sartori</i>	
Looking for Efficient Methods to Collect and Geolocalise Tweets	1057
Alla ricerca di metodi efficienti per raccogliere e geolocalizzare tweet	
<i>Stephan Schlosser, Daniele Toninelli and Silvia Fabris</i>	
Principal ranking profiles	1063
Principal ranking profiles	
<i>Mariangela Sciandra, Antonella Plaia</i>	
A statistical model for voting probabilities	1069
Un modello statistico per le probabilità di voto	
<i>Rosaria Simone, Stefania Capecci</i>	
How Citizen Science and smartphones can help to produce timely and reliable information? Evidence from the "Food Price Crowdsourcing in Africa" (FPCA) project in Nigeria	1075
Citizen Science e smartphone posso aiutare nella raccolta di dati tempestivi e affidabili? Testimonianze del progetto "Food Price Crowdsourcing in Africa" (FPCA) condotto in Nigeria	
<i>Gloria Solano-Hermosilla, Fabio Micale, Vincenzo Nardelli, Julius Adewopo, Celso Garrín González</i>	
Dealing with uncertainty in automated test assembly problems	1083
La gestione dell'incertezza nei problemi di assemblaggio automatizzato dei test	
<i>Giada Spaccapanico Proietti, Mariagiulia Matteucci and Stefania Mignani</i>	
Joint Models: a smart way to include functional data in healthcare analytics	1089
Modelli congiunti: un metodo per includere i dati funzionali nelle analisi in ambito sanitario	
<i>Marta Spreafico, Francesca Ieva</i>	
Bayesian multiscale mixture of Gaussian kernels for density estimation	1095
Stima di densità tramite misture bayesiane multiscala di kernel gaussiani	
<i>Marco Stefanucci and Antonio Canale</i>	
Dynamic Bayesian clustering of running activities	1101
Clustering Bayesiano dinamico di attività di corsa	
<i>Mattia Sival and Mauro Bernardi</i>	
Employment and fertility in couples: whose employment uncertainty matter most?	1107
Lavoro e fecondità in coppia: il ruolo dell'incertezza lavorativa secondo una prospettiva di genere	
<i>Valentina Tocchioni, Daniele Vignoli, Alessandra Mattei, Bruno Arpino</i>	
A Functional Data Analysis Approach to Study a Bike Sharing Mobility Network in the City of Milan	1113
<i>Agostino Torti, Alessia Pini and Simone Vantini</i>	
Multiresolution Topological Data Analysis for Robust Activity Tracking	1119
<i>Giovanni Trappolini, Tullia Padellini, and Pierpaolo Brutti</i>	
Semilinear regression trees	1125
Alberi di regressione semilineari	
<i>Giulia Vannucci and Anna Gottard</i>	

A models selection criterion for evaluation of heat wave hazard: a case study of the city of Prato.....	1131
Un criterio di selezione dei modelli per la valutazione della pericolosità delle ondate di calore: un caso studio della città di Prato	
<i>Veronica Villani, Giuliana Barbato, Elvira Romano and Paola Mercogliano</i>	
Digital Inequalities and ICT Devices: The ambiguous Role of Smartphones.....	1139
<i>Laura Zannella, Marina Zannella</i>	

Section 4. Posters

Modelling Hedonic Price using semiparametric M-quantile regression	1147
Regressione m-quantilica semiparametrica per la modellizzazione dei prezzi edonici	
<i>Riccardo Borgoni, Antonella Carcagni, Alessandra Michelangeli, Nicola Salvati</i>	
Bayesian mixed latent factor model for multi-response marine litter data with multi-source auxiliary information	1153
Modello bayesiano misto a fattori latenti per l'abbondanza di rifiuti marini con informazioni ausiliarie di diversa provenienza	
<i>Crescenza Calcutti, Alessio Pollice, Marco V. Guglielmi and Porzia Maiorano</i>	
Official statistics to support the projects of A Scuola di OpenCoesione	1159
L'esperienza di monitoraggio civico in Lombardia nell'anno scolastico 2018-19	
<i>del Vicario G. and Di Gennaro L. and Ferrazza D. and Spinella V. and Viviano L.</i>	
Spatial Logistic Regression for Events Lying on a Network: Car Crashes in Milan	1165
Regressione logistica per eventi su network: gli incidenti automobilistici nel comune di Milano	
<i>Andrea Gilardi, Riccardo Borgoni and Diego Zappa</i>	
Variable selection and classification by the GRID procedure	1171
Selezione e classificazione delle variabili attraverso il metodo GRID	
<i>Francesco Giordano, Soumendra Nath Lahiri and Maria Lucia Parrella</i>	
Joint VaR and ES forecasting in a multiple quantile regression framework.....	1177
Stima congiunta del VaR e dell'ES attraverso la regressione quantilica multipla	
<i>Merlo Luca, Petrella Lea and Raponi Valentina</i>	
Approximate Bayesian Computation methods to model Multistage Carcinogenesis	1183
Metodi di Approximate Bayesian Computation per modellare la Cancerogenesi Multistadiale	
<i>Consuelo R. Nava, Cinzia Carota, Jordy Bollon, Corrado Magnani, Francesco Barone-Adesi</i>	
Co-clustering TripAdvisor data for personalized recommendations	1189
Co-clustering di dati TripAdvisor per un sistema di raccomandazioni personalizzato	
<i>Giulia Pascali, Alessandro Casa and Giovanna Menardi</i>	
Latent class analysis of endoreduplicated nuclei in confocal microscopy.....	1195
Analisi di classi latenti per dati di nuclei endoreduplicati tramite microscopia confocale	
<i>Ivan Sciascia ivan.sciascia@unito.it, Gennaro Carotenuto gennaro.carotenuto@unito.it, Andrea Genre andrea.genre@unito.it, Università di Torino Dipartimento di Scienze della vita e biologia dei sistemi, viale Mattioli 25, 10125 Torino</i>	

Design of Experiments, aberration and Market Basket Analysis

Pianificazione degli esperimenti, aberrazione e Market Basket Analysis

Roberto Fontana and Fabio Rapallo

Abstract In Design of Experiments, the Minimum Aberration criterion uses the aberrations of all the main effects and interactions to choose *suitable* fractions. In this work, we consider the aberrations in Market Basket Analysis, which is a completely different context from Design of Experiments. Using a real dataset, we show how aberrations could represent a meaningful measure of association.

Abstract *Nella pianificazione degli esperimenti, il criterio della Minima Aberrazione utilizza le aberrazioni di tutti gli effetti principali e di interazione per selezionare opportuni piani fattoriali frazionari. In questo contributo consideriamo le aberrazioni nel contesto – totalmente diverso – della Market Basket Analysis. Basandoci su un esempio applicativo, mostriamo che le aberrazioni rappresentano una buona misura di associazione.*

Key words: Fractional factorial designs, Word-Length Pattern, association

1 Introduction

Fractional factorial designs are commonly used in Design of Experiments to determine the optimum mix of factors to predict a response variable. The need for efficient experimental designs has led to the definition of several criteria for the choice of the design points. In the case of binary designs, an important object associated to a design is its Word-Length Pattern (WLP). The WLP is used to discriminate among different designs through the Minimum Aberration (MA) criterion, which is based on the sequential minimization of the WLP. The MA criterion was introduced in

Roberto Fontana
DISMA, Dipartimento di Eccellenza 2018-2022, Politecnico di Torino, e-mail:
roberto.fontana@polito.it

Rapallo Fabio
DISIT, Università del Piemonte Orientale e-mail: fabio.rapallo@uniupo.it

[6] for binary designs and then extended with the name of Generalized Minimum Aberration to non-regular multilevel designs in [7]. The WLP is computed using the aberrations of all the main effects and interactions.

In this work we use aberrations in the context of Market Basket Analysis (MBA). MBA finds association rules between sets of products bought in a unit of shopping by consumers (see e.g. [1]). In our application, given m products, a basket will be a binary vector $x = (x_1, \dots, x_m)$, where $x_i = 1$ (resp. $x_i = -1$) means that i -th product has (resp. has not) been bought by the customer. The set of all the baskets is a multiset of $\{-1, 1\}^m$ and can be considered as a fractional factorial design of $\{-1, 1\}^m$. We use aberrations to describe the associations between products. To the best of our knowledge the use of aberrations as measures of association is new. An application to a set of Italian museums is shown.

2 Fractional factorial designs and aberrations

Let us consider an experiment with m binary factors. The full factorial design is $\mathcal{D} = \{-1, 1\}^m$. We briefly recall here the basic definitions concerning fractional factorial designs (or simply fractions) and aberrations. For details refer to [4].

Definition 1. A fraction \mathcal{F} is a multiset (\mathcal{F}_*, f_*) whose underlying set of elements \mathcal{F}_* is contained in \mathcal{D} and f_* is the multiplicity function $f_* : \mathcal{F}_* \rightarrow \mathbb{N}$ that for each element in \mathcal{F}_* gives the number of times it belongs to the multiset \mathcal{F} .

We recall that the underlying set of elements \mathcal{F}_* is the subset of \mathcal{D} that contains all the elements of \mathcal{D} that appear in \mathcal{F} at least once. We denote the number of elements of the fraction \mathcal{F} by $\#\mathcal{F}$, with $\#\mathcal{F} = \sum_{x \in \mathcal{F}_*} f_*(x)$.

To describe the counting function of a fraction, we follow the theory in [4]. The simple terms of the form X_j , i.e., the j -th component function which maps a point $x = (x_1, \dots, x_m)$ of \mathcal{D} to its j -th component,

$$X_j : \mathcal{D} \ni (x_1, \dots, x_m) \mapsto x_j \in \{-1, 1\}$$

and the interactions $X^\alpha = X_1^{\alpha_1} \cdot \dots \cdot X_m^{\alpha_m}$, $\alpha \in L = \{0, 1\}^m$ i.e., the monomial functions of the form

$$X^\alpha : \mathcal{D} \ni (x_1, \dots, x_m) \mapsto x_1^{\alpha_1} \cdot \dots \cdot x_m^{\alpha_m}$$

are a basis of all the real functions defined over \mathcal{D} . We use this basis to represent the counting function of a fraction according to the following definition.

Definition 2. The counting function R of a fraction \mathcal{F} is a polynomial defined over \mathcal{D} so that for each $x \in \mathcal{D}$, $R(x)$ equals the number of appearances of x in the fraction. We denote by c_α the coefficients of the representation of R on \mathcal{D} using the monomial basis $\{X^\alpha, \alpha \in L\}$:

$$R(x) = \sum_{\alpha \in L} c_\alpha X^\alpha(x), \quad x \in \{-1, 1\}^m, \quad c_\alpha \in \mathbb{R}.$$

Definition 3. The Word-Length Pattern (WLP) of a fraction \mathcal{F} of the full factorial design \mathcal{D} is the vector $A_{\mathcal{F}} = (A_0(\mathcal{F}), A_1(\mathcal{F}), \dots, A_m(\mathcal{F}))$, where

$$A_j(\mathcal{F}) = \sum_{|\alpha|_0=j} a_{\alpha} \quad j = 0, \dots, m,$$

$$a_{\alpha} = (c_{\alpha}/c_0)^2,$$

$|\alpha|_0$ is the number of non-null elements of α , and $c_0 := c_{(0, \dots, 0)} = \#\mathcal{F}/\#\mathcal{D}$.

In the definition above, the number a_{α} is the aberration of the term X^{α} . In the case of binary design, using some results in [5], Prop. 1 below yields the aberration a_{α} as a function of the counts of the corresponding term X^{α} .

Proposition 1. Given a fraction \mathcal{F} of $\mathcal{D} = \{-1, 1\}^m$ the aberration a_{α} of the term X^{α} is

$$a_{\alpha} = (n_{1,\alpha} - n_{-1,\alpha})^2 / \#\mathcal{F}^2, \tag{1}$$

where $n_{1,\alpha}$ (resp. $n_{-1,\alpha}$) is the number of times X^{α} is equal 1 (resp. -1) in \mathcal{F} ,

$$n_{1,\alpha} = \sum_{x \in \mathcal{D}} R(x) \delta_x^{1,\alpha}, \quad n_{-1,\alpha} = \sum_{x \in \mathcal{D}} R(x) \delta_x^{-1,\alpha},$$

and δ_a^b denotes the Kronecker delta.

It is worth noting that the terms a_{α} can be easily interpreted for main effects and 2-factor interactions, i.e. when $|\alpha|_0 \in \{1, 2\}$. If $|\alpha|_0 = 1$, without loss of generality, let us consider $\alpha = (1, 0, \dots, 0)$. We get:

$$n_{1,\alpha} \equiv n_{1,(1,0,\dots,0)} = \sum_{x=(x_1,\dots,x_m) \in \mathcal{D}: x_1=1} R(x) \tag{2}$$

and a similar formula holds for $n_{-1,\alpha} \equiv n_{-1,(1,0,\dots,0)}$. It follows that a_{α} is the squared difference between the relative frequency of the points for which $x_1 = 1$ and the relative frequency of the points for which $x_1 = -1$.

If $|\alpha|_0 = 2$ let us consider $\alpha = (1, 1, 0, \dots, 0)$. We get:

$$n_{1,\alpha} \equiv n_{1,(1,1,0,\dots,0)} = \sum_{\substack{x=(x_1,\dots,x_m) \in \mathcal{D}: \\ x_1 x_2 = 1}} R(x) = \sum_{\substack{x=(x_1,\dots,x_m) \in \mathcal{D}: \\ x_1 = x_2 = 1}} R(x) + \sum_{\substack{x=(x_1,\dots,x_m) \in \mathcal{D}: \\ x_1 = x_2 = -1}} R(x) \tag{3}$$

and a similar formula holds for $n_{-1,\alpha} \equiv n_{-1,(1,1,0,\dots,0)}$. In this case a_{α} is the squared difference between the relative frequency of the points for which there is *agreement*, i.e. $x_1 = x_2 = 1$ or $x_1 = x_2 = -1$ and the relative frequency of the points for which there is *disagreement*, i.e. $x_1 = -1, x_2 = 1$ or $x_1 = 1, x_2 = -1$.

To correctly interpret the aberrations of order two, we compute the approximate distributions of such parameters through a Monte Carlo resampling (based on 1,000 replicates) for different values of the sample size, in order to check also the robustness of the method for small data sets. Moreover, we compare the approximate distribution of each aberration with the expected aberration under independence.

For instance, the expected value under independence of the aberration $a_{(1,1,0,\dots,0)}$ is:

$$\begin{aligned} \widehat{a}_{(1,1,0,\dots,0)} &= (\widehat{n}_{1,(1,1,0,\dots,0)} - \widehat{n}_{-1,(1,1,\dots,0)})^2 / \#\mathcal{F}^2 = \\ &= \{n_{1,(1,0,0,\dots,0)}(n_{1,(0,1,0,\dots,0)} - n_{-1,(0,1,0,\dots,0)}) \\ &\quad - n_{-1,(1,0,0,\dots,0)}(n_{1,(0,1,0,\dots,0)} - n_{-1,(0,1,0,\dots,0)})\}^2 / \#\mathcal{F}^4. \end{aligned}$$

By comparison with Eqs. (1) and (2), we get:

$$\widehat{a}_{(1,1,0,\dots,0)} = a_{(1,0,0,\dots,0)} a_{(0,1,0,\dots,0)}. \quad (4)$$

3 An application to Italian museums

In this paper we study a subset of museums participating in the Abbonamento Musei Torino Piemonte (AMTP, <https://piemonte.abbonamentomusei.it>) network. This network was created in 1995 and is available to people living in the Piemonte region (Italy). For a yearly subscription fee, AMTP card-holders have free entry to all the museums and all the temporary/permanent exhibitions participating in the program for the subscription year, from January to December. In recent years the number of subscribers has increased enormously: from 5,734 cards in 1999 to 87,237 cards in 2012, to 127,768 in 2016.

We analyze the 2012 AMTP transaction database. A descriptive analysis of the association structure of the 2012 AMTP network was conducted in [2]. Graphical models have been studied in [3]. The 2012 AMTP database collects information for each card-holder about the museums that he/she visited in 2012.

We consider which museums have been visited by each card-holder, regardless of the number of times he/she returned to the same museum. In this way we exclude repeated visits and we have a total 287,259 visits to the main 23 museums. The number of card-holders who visited at least one of the 23 museums is 73,668. Therefore, the final dataset has $N = 73,668$ observations and is organized as follows. Each row corresponds to a card-holder. The first 23 columns are binary variables and indicate whether the card-holder visited a museum one or more times, or did not visit a museum. As explained above we do not distinguish between single and multiple visits to the same museum. We note that the number of observations ($N = 73,668$) is less than the number of card-holders in 2012 because some of them did not use the card or did not visit any of the 23 museums which have been chosen for this analysis. Given the limited amount of space for this paper we select five museums, build up a $\{-1, 1\}^5$ design with an appropriate counting function containing the absolute frequencies of visits, and show the use of aberrations of order two as one of the statistics to describe associations between any two of them.

In our study a *basket* is the set of museums which have been visited at least one time in the same year by an individual card-holder.

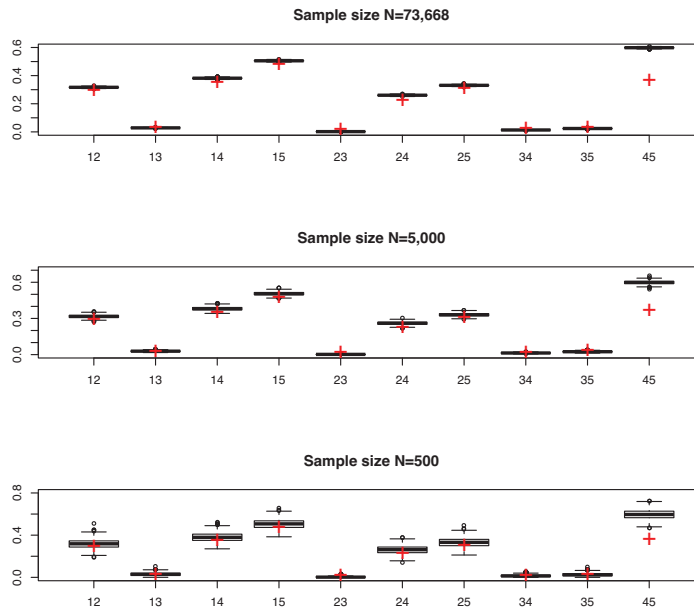


Fig. 1 Distribution of the aberrations of order two for different sample sizes. The symbol ‘+’ denotes the expected aberration under independence.

In Fig. 1 the distributions (boxplots) of the aberrations of order two are compared with the expected aberration as in Eq. (4) for three sample sizes, namely the original sample size $N = 73,668$, $N = 5,000$ and $N = 500$. We note that the aberration 45 is significantly higher than expected, showing an agreement between the last two museums. We can also observe that the analysis remains robust when the sample size decreases.

We also compare the aberrations with the classical correlation coefficients. The results are displayed in Fig. 2. Here we note that there is a moderate correlation between the second and the third museum, not recognized by the aberration. This fact happens because the third museum has a large number of visitors compared with the other museums. Indeed, we can observe the same behavior for all the aberrations involving the third museum. In this example, the aberrations are less sensitive than correlations to non-homogeneous margins, and this may be an issue in favor of aberrations, especially when some museums has few visitors.

Future research will focus on the use of aberrations of order greater than two to measure the associations of more than two variables. The comparison of aberrations with other common measures of association (e.g. odds ratios and lifts) will also be studied.

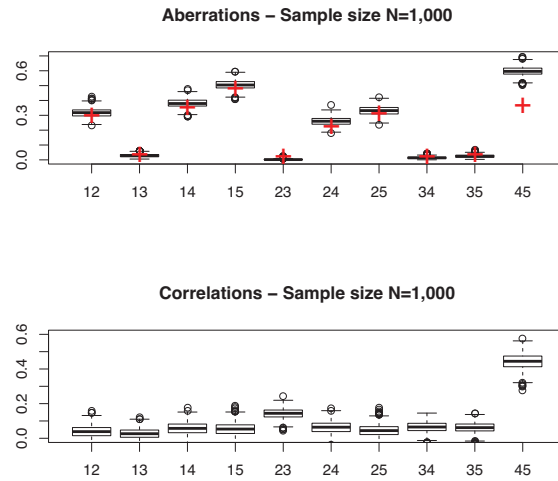


Fig. 2 Distribution of the aberrations of order two and the corresponding correlation coefficients.

Acknowledgements R. Fontana acknowledges that the present research has been partially supported by MIUR grant Dipartimenti di Eccellenza 2018-2022 (E11G18000350001). The authors thank Francesca Leon, past president of the Associazione Torino Città Capitale and currently member of the Turin city council, for providing the database of 2012 AMTP card-holders. R. Fontana also thanks Prof. Patrizia Semeraro (Politecnico di Torino) for helpful discussions.

References

1. Agresti, A.: *Categorical data analysis*, 3rd edn. John Wiley & Sons (2013)
2. Coscia, C., Fontana, R., Semeraro, P.: Market basket analysis for studying cultural consumer behaviour: AMTP card-holders. *Statistica Applicata - Italian Journal of Applied Statistics* **26**(2), 73–92 (2016)
3. Coscia, C., Fontana, R., Semeraro, P.: Graphical models for complex networks: an application to Italian museums. *J. Appl. Stat.* **45**(11), 2020–2038 (2018)
4. Fontana, R., Pistone, G., Rogantin, M.P.: Classification of two-level factorial fractions. *J. Statist. Plann. Inference* **87**(1), 149–172 (2000)
5. Fontana, R., Rapallo, F., Rogantin, M.P.: Aberration in qualitative multilevel designs. *J. Statist. Plann. Inference* **174**, 1–10 (2016)
6. Fries, A., Hunter, W.G.: Minimum aberration 2^{k-p} designs. *Technometrics* **22**(4), 601–608 (1980)
7. Xu, H., Wu, C.F.J.: Generalized minimum aberration for asymmetrical fractional factorial designs. *Ann. Statist.* **29**(4), 1066–1077 (2001)