

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 Alburez-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 Lupparelli</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 Neurali 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.	
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., Scepti 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 Marialuca 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 Diqiugiovanni</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 multivariati per serie storiche climatiche</i>	
<i>Gianluca Mastrantonio, Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio and Carlo Blasi</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 Mameli 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'Ambrà, D'Ambrà 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 Schellin, 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 Schuberth</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 Fucella, 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
Misure 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 Sinitò, 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 basate 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., Guatteri, 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 Zagariou</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
<i>Scommettere sul calcio: un nuovo modello per prevedere l'esito delle partite</i>	
<i>Marco Petretta, Lorenzo Schiavon and Jacopo Diquigiovanni</i>	
Estimation of dynamic quantile models via the MM algorithm	1033
<i>Stima di modelli Quantilici Dinamici con algoritmo MM</i>	
<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
<i>La scomposizione per sottopopolazioni dell'indice di Pietra: un'applicazione alle squadre professionistiche di calcio in Italia</i>	
<i>Francesco Porro and Mariangela Zenga</i>	
An Object Oriented Data Analysis of Tweets: the Case of Queen Elizabeth Olympic Park .	1045
<i>Object Oriented Data Analysis di Tweet: il caso del Queen Elizabeth Olympic Park</i>	
<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
<i>Stima non distorta di un modello Expected Goal con effetti fissi nel calcio</i>	
<i>Lorenzo Schiavon and Nicola Sartori</i>	
Looking for Efficient Methods to Collect and Geolocalise Tweets	1057
<i>Alla ricerca di metodi efficienti per raccogliere e geolocalizzare tweet</i>	
<i>Stephan Schlosser, Daniele Toninelli and Silvia Fabris</i>	
Principal ranking profiles	1063
<i>Principal ranking profiles</i>	
<i>Mariangela Sciandra, Antonella Plaia</i>	
A statistical model for voting probabilities	1069
<i>Un modello statistico per le probabilità di voto</i>	
<i>Rosaria Simone, Stefania Capecchi</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
<i>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>	
<i>Gloria Solano-Hermosilla, Fabio Micale, Vincenzo Nardelli, Julius Adewopo, Celso Gorrín González</i>	
Dealing with uncertainty in automated test assembly problems	1083
<i>La gestione dell'incertezza nei problemi di assemblaggio automatizzato dei test</i>	
<i>Giada Spaccapanico Proietti, Mariagiulia Matteucci and Stefania Mignani</i>	
Joint Models: a smart way to include functional data in healthcare analytics	1089
<i>Modelli congiunti: un metodo per includere i dati funzionali nelle analisi in ambito sanitario</i>	
<i>Marta Spreafico, Francesca Ieva</i>	
Bayesian multiscale mixture of Gaussian kernels for density estimation	1095
<i>Stima di densità tramite misture bayesiane multiscala di kernel gaussiani</i>	
<i>Marco Stefanucci and Antonio Canale</i>	
Dynamic Bayesian clustering of running activities	1101
<i>Clustering Bayesiano dinamico di attività di corsa</i>	
<i>Mattia Stival and Mauro Bernardi</i>	
Employment and fertility in couples: whose employment uncertainty matter most?	1107
<i>Lavoro e fecondità in coppia: il ruolo dell'incertezza lavorativa secondo una prospettiva di genere</i>	
<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
<i>Alberi di regressione semilineari</i>	
<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 Calculli, 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>	



Preface

Preface

This book includes the scientific contributions presented at the Intermediate Meeting of Italian Statistical Society (SIS) held in Milan at the Università Cattolica del Sacro Cuore, from June 18th to 21th of 2019. Following a long tradition (and a statutory indication of the Society), the intermediate meetings are held bi-annually on specific themes. This year, aiming at bridging the gap between statistics and the world of Big Data and Data Science, the conference was entirely devoted to the theme of “Smart Statistics for Smart Applications”. In this way the Italian Statistical Society had the explicit intention to answer the high and rapidly increasing demand on the subject, by providing academics, researchers and practitioners with a forum where new ideas and new methods could meet with new needs, new research questions and new applications.

The Conference could not have been organized without the joint effort of the Milanese network of Università Cattolica del Sacro Cuore, Università degli Studi di Milano Bicocca, Università Bocconi, Università “Vita e salute” San Raffaele, Politecnico di Milano and Università Statale di Milano. Members of all these universities took part actively to the Local Organizing Committee. The Conference has also greatly benefited from the contribution of the strategic partner Mathesia, which contributed to the various aspects of the organization, with special focus to the active involvement of private firms and companies and of the non-academic components.

The conference has registered more than 200 scientific contributions, including papers presented in plenary invited sessions, papers collected in specialized and solicited sessions on specific themes, about 100 contributions spontaneously submitted to the Program Committee and a poster session. All contributions were focused on the conference theme and provided a good overview of the state-of-the-art of the subject, from methodological and theoretical contributions, to applied works and case studies. The two plenary lectures were devoted to the (provocative) idea of “shallow learning”, as opposed to the more in-vogue idea of deep learning”, and to the problems linked with Big Data veridicity and reliability. A plenary round table draw the participants attention on the concept of smart ageing.

A distinctive feature of this conference, relative to previous analogous experiences, was the presence of many round tables and activities focused on topics of interest for a wider audience, freely open to external participation. These activities were termed “Fuoricongresso” and included a special session on “Data skills: Statistics and education for future jobs” organized jointly with Pearson Italia Publisher, a round table on “How to Close the Gap Between the Practice and Theory in Digital Transformation Era” organized joint with Mathesia, a colloquium on “Big Data and Big Responsibility”, a round table on “Political polls in the Big Data era”, a round table on “Big Data and Public Administration”, a round table on the changing role of the statistical scientific societies in a new interconnected world and the fifth edition of the statistical competition “Stats Under the Stars (SUS5)” organized by the Bocconi University, a whole-night hackathon on real-world business analytic problems for young Data Scientists.

More information about the fuoricongresso activities may be found on the website of the meeting¹. We offer this book to all members of the Italian Statistical Society, to all participants of the conference and to all interested people, in the hope that this will provide them with a good snapshot of the on-going research in this exciting new area of statistical studies. We deeply thank all contributors for having submitted their work to the conference and all the researchers who did an outstanding job in acting as referees accurately and timely. Finally we wish to express our gratitude to the publisher Pearson Italia for all the support received.

Giuseppe Arbia
Stefano Peluso
Alessia Pini
Giulia Rivellini

¹ URL: <http://meetings3.sis-statistica.org/index.php/SIS2019/sis2019/schedConf/overview>

Predicting and improving smart mobility: a robust model-based approach to the BikeMi BSS

Prevedere e migliorare la mobilità smart: un approccio robusto di classificazione applicato a BikeMi

Andrea Cappelozzo, Francesca Greselin and Giancarlo Manzi

Abstract Bike Sharing Systems play a central role in what is identified to be one of the six pillars of a Smart City: smart mobility. Motivated by a freely available dataset, we discuss the employment of two robust model-based classifiers for predicting the occurrence of situations in which a bike station is either empty or full, thus possibly creating demand loss and customer dissatisfaction. Experiments on BikeMi stations located in the central area of Milan are provided to underline the benefits of the proposed methods.

Abstract *I sistemi di Bike Sharing giocano un ruolo centrale nella mobilità sostenibile, uno dei sei pilastri che indentificano una Smart City. Motivati da un set di dati disponibile online, questo lavoro presenta l'utilizzo di due modelli di classificazione robusta per prevedere il manifestarsi di situazioni in cui una bike station sia piena e/o vuota, così creando perdita di domanda ed insoddisfazione nei clienti. Esperimenti di classificazione sulle stazioni BikeMi nel centro di Milano evidenziano l'efficacia dei metodi proposti.*

Key words: Bike Sharing System, Smart Mobility, Impartial Trimming, Robust Classification

1 Motivating problem

The world's population forecast is estimated to reach 9 billions in the upcoming years, with up to 66% of the total humankind living in urbanized areas [6]. The

Andrea Cappelozzo • Francesca Greselin
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, e-mail: a.cappelozzo@campus.unimib.it; francesca.greselin@unimib.it

Giancarlo Manzi
Department of Economics, Management and Quantitative Methods, University of Milan, e-mail: giancarlo.manzi@unimi.it

urban ecosystem devoted to accommodate such a huge proportion of the future population will most likely be a *smart city*: a new metropolitan vision that integrates information and communications technology (ICT) and physical infrastructure, encompassing every municipality aspect: from mobility to architecture, infrastructure and power supply management [7]. Particularly, six pillars identify and assess the concept of “smartness” in such context: economy, people, governance, environment, living and mobility [12].

The present work will focus on sustainable mobility and specifically on the analysis of the BikeMi bike sharing system (BSS) in Milan, as an environmental friendly complement to public and private transports, with the final aim of assessing and possibly improving the service. It is well known that BSS with docking stations users identify finding an available bicycle and a parking slot as the two most critical problems in their biking experience [4]. By employing a robust classification model we try to predict whether and when these problems might occur, identifying some useful insights that may be of use in subsequently planning manual bicycle repositioning.

The rest of the manuscript is organized as follows: in Section 2 the main characteristics of the BikeMi bike sharing system is described; together with the dataset considered in the study. Section 3 details the robust classification method employed in predicting possible future FULL/EMPTY stations scenarios, with the analysis results presented in Section 4. The paper concludes with a list of proposals for future research direction.

2 The BikeMi BSS

BikeMi was introduced in November 2008 as the first privately managed Italian bike sharing system [9]. Presently, the service encompasses 280 active stations for a total of 4650 available bikes. The dataset considered in this manuscript reports the stations status, in terms of available bikes and free slots, during the period January-August 2015. Records were periodically collected by scraping the BikeMi website: the full dataset is publicly available online [11]. The average weekday profile usage in terms of Normalized Available Bikes (number of bikes / total number of slots in the station) is represented in Figure 1. From the plot two main distinct behaviors are visible: stations that are almost full in the morning and get gradually empty, and stations that follow a mirror pattern. Such a scheme is primarily driven by morning and evening work commuters; it is therefore essential for the BSS success to be able to efficiently cope with this daily bikes demand.

The aim of the present work is to develop a classifier that will help in predicting whether a demand loss might occur. Specifically, since lost demand arises as a consequence of full or empty station, we build a method that predicts how likely are such situations to happen given a set of available information. Time features, past inventory features and meteorological variables are employed in building the classification rule. Given the noisy nature of the dataset at hand (i.e., stations and slots are prone to malfunctions and breakage, undermining the quality of the scrap-

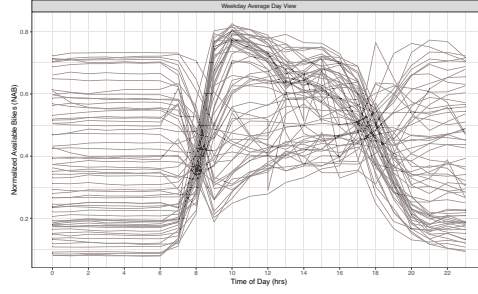


Fig. 1 Average weekday Normalized Available Bikes (number of bikes / total number of slots in the station) for the BikeMi stations in Milan central area.

ing) we propose to employ two robust model-based classifiers for determining the future FULL, EMPTY or NOT PROBLEMATIC status of a particular station. The employed methods are described in the next Section.

3 Robust model-based classifiers

Let $\{(\mathbf{x}_1, \mathbf{l}_1), \dots, (\mathbf{x}_N, \mathbf{l}_N)\}$ be a complete set of learning observations, where \mathbf{x}_n denotes a p -variate observation and \mathbf{l}_n its associated class label, such that $l_{ng} = 1$ if observation n belongs to group g and 0 otherwise, $g = 1, \dots, G$, $n = 1, \dots, N$. In the context of our analysis we set $G = 3$, to define the FULL, EMPTY or NOT PROBLEMATIC status of a station in the future time-slot. Likewise, denote the set of unlabelled observations by \mathbf{y}_m , $m = 1, \dots, M$ and their associated unknown labels z_{mg} , $g = 1 \dots G$ and $m = 1, \dots, M$. We construct a procedure for maximizing the *trimmed observed data log-likelihood*:

$$\begin{aligned} \ell_{trim}(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{X}, \mathbf{Y}, 1) = & \sum_{n=1}^N \zeta(\mathbf{x}_n) \sum_{g=1}^G l_{ng} \log [\tau_g \phi(\mathbf{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)] + \\ & + \sum_{m=1}^M \varphi(\mathbf{y}_m) \log \left[\sum_{g=1}^G \tau_g \phi(\mathbf{y}_m; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right] \end{aligned} \quad (1)$$

where τ_g is the prior probability of observing class g ; $\phi(\cdot; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the multivariate normal density with mean vector $\boldsymbol{\mu}_g$ and variance covariance matrix $\boldsymbol{\Sigma}_g$; $\zeta(\cdot)$ and $\varphi(\cdot)$ are 0-1 trimming indicator functions, that express whether observation \mathbf{x}_n and \mathbf{y}_m are trimmed off or not. The *labelled trimming level* α_l , s.t. $\sum_{n=1}^N \zeta(\mathbf{x}_n) = \lceil N(1 - \alpha_l) \rceil$ and the *unlabelled trimming level* α_u , s.t. $\sum_{m=1}^M \varphi(\mathbf{y}_m) = \lceil M(1 - \alpha_u) \rceil$ account for possible noisy observations and outliers in both sets.

The aforementioned specification leads to two robust model-based classification approaches: if only the labelled observations are employed for estimating parameters (i.e., only the first line of (1) is considered) we obtain a Robust Eigenvalue

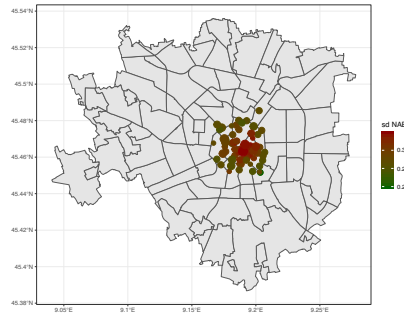


Fig. 2 Location of the BikeMi stations considered in the analysis. Dots size denotes the station total number of slots, the color scaling indicates the standard deviation of normalized bikes availability averaged per weekday.

Decomposition Discriminant Analysis (REDDA); whereas we retrieve a Robust Updating Classification Rule (RUPCLASS) if a semi-supervised approach is favoured. These models are robust generalizations of the techniques developed in [1] and [2], respectively.

Parameters estimation is carried out via a procedure similar to the FastMCD algorithm [8] for the REDDA model, and via the EM algorithm [3] with an appropriate Concentration Step [8] and eigenvalue-ratio restriction [5] enforced at each iteration for RUPCLASS.

4 Classification results

The methodologies described in the previous Section are employed for predicting the station status one hour in the future, thus assessing the need of manual bikes repositioning when FULL and EMPTY situations are forecast. The stations in the analysis are a subset of the ones located in the central area of the city (Bastioni and Centro Storico): a spatial representation is reported in Figure 2.

The considered time-frame is limited to the quarter April-June 2015, in which the last eight days of June are kept out from the learning set and used for assessing the prediction accuracy. The classification results for the models described in Section 3 are reported in Table 1. The classification rates is on average above 0.82 for both REDDA and RUPCLASS, even if the supervised model seems to perform slightly better overall. Particularly, this is more predominant for stations that present a higher turnover, where the unlabelled set provides less useful information about separation between groups [10].

Title Suppressed Due to Excessive Length

Table 1 Correct classification rates for the BikeMi stations in Milan central area for the last eight days of June 2015 (22-30) employing REDDA and RUPCLASS models.

Station ID	Station Name	REDDA	RUPCLASS
1	Duomo	0.719	0.733
3	Cadorna 1	0.641	0.613
4	Lanza	0.793	0.797
5	Università Cattolica	0.843	0.770
20	Erculea	0.853	0.894
34	Cairoli	0.806	0.788
37	Italia - San Martino	0.931	0.945
43	Festa del Perdono	0.811	0.774
44	Richini	0.912	0.894
45	Cant	0.820	0.811
54	Sant'Eustorgio- P.ta Ticinese	0.853	0.811
60	Edison	0.839	0.816
63	Sant'Ambrogio	0.788	0.806
64	Diaz	0.816	0.811
84	Cadorna 2	0.806	0.848
94	Cadorna 3	0.673	0.668
13	Senato	0.728	0.843
14	San Barnaba H Mangiagalli	0.871	0.912
15	Cantore	0.912	0.912
16	Moscova	0.774	0.779
22	Medaglie D'Oro 1	0.779	0.857
23	Regina Margherita	0.922	0.935
25	Centrale 1	0.687	0.673
27	Porta Venezia	0.797	0.802
30	Crocetta	0.848	0.857
32	Manin - Bastioni	0.880	0.908
46	Porta Nuova	0.876	0.848
55	Cinque Giornate	0.871	0.899
58	Sant'Agostino	0.908	0.889
88	Beatrice d'Este - Cassolo	0.945	0.959
98	San Marco	0.945	0.908
99	Arco della Pace 1 - Bertani	0.903	0.894
103	Arco della Pace 2 - Pagano	0.899	0.889
181	Sempione - Melzi d Eril	0.917	0.922

The present work employs two robust model-based classifiers for detecting possible situations of future demand loss for the BikeMi BSS in Milan. The classification accuracy obtained for a subset of stations in the central area fosters the employment of the described methods. Further research directions will consider the integration of spatial information related to the inventory of the stations closest to the target, and the employment of a cost function for over-penalizing the misclassification of FULL and EMPTY statuses as NOT PROBLEMATIC.

References

1. H. Bensmail and G. Celeux. Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91(436):1743–1748, dec 1996.
2. N. Dean, T. B. Murphy, and G. Downey. Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 55(1):1–14, 2006.
3. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.
4. J. Froehlich, J. Neumann, and N. Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *IJCAI International Joint Conference on Artificial Intelligence*, 2009.
5. S. Ingrassia. A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods and Applications*, 13(2):151–166, 2004.
6. S. Mallapuram, N. Ngwum, F. Yuan, C. Lu, and W. Yu. Smart city: The state of the art, datasets, and evaluation platforms. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pages 447–452. IEEE, may 2017.
7. I.-I. Picioroaga, M. Eremia, and M. Sanduleac. SMART CITY: Definition and Evaluation of Key Performance Indicators. In *2018 International Conference and Exposition on Electrical And Power Engineering (EPE)*, pages 217–222. IEEE, oct 2018.
8. P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, aug 1999.
9. G. Saibene and G. Manzi. Bike usage in public bike-sharing: An analysis of the BikeMi system in Milan. Technical report, 2015.
10. D. Toher, G. Downey, and T. B. Murphy. A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies. *Chemometrics and Intelligent Laboratory Systems*, 89(2):102–115, nov 2007.
11. A. Trentini. Scraped Data BikeMI 2015. <http://doi.org/10.5281/zenodo.1209270>, mar 2018.
12. J. Zawieska and J. Pieriegud. Smart city as a tool for sustainable mobility and transport decarbonisation. *Transport Policy*, 63:39–50, apr 2018.