



Università  
Bocconi  
MILANO



POLITECNICO  
MILANO 1863



UNIVERSITÀ DEGLI STUDI DI MILANO  
UNIVERSITÀ DEGLI STUDI DI MILANO



# 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 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 Neurali nella Previsione di Serie Storiche	
<i>Giuseppe Bruno, Sabina Marchetti, Juri Marcucci, Diana Nicoletti</i>	

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

<b>A proposal for an integrated approach between sentiment analysis and social network analysis.....</b>	<b>247</b>
Una proposta per un approccio integrato tra analisi del sentimento e analisi delle reti sociali	
<i>Domenico De Stefano and Francesco Santelli</i>	
<b>A meta-tissue non-parametric factor analysis model for gene co-expression .....</b>	<b>255</b>
Meta-analisi fattoriale non parametrica per lo studio di espressioni genetiche in diversi tessuti	
<i>Roberta De Vito and Barbara Engelhardt</i>	
<b>Bayesian estimate of population count with false captures: a latent class approach.....</b>	<b>261</b>
Stima Bayesiana della popolazione con false catture: un approccio basato sulle classi latenti	
<i>Davide Di Cecco, Marco Di Zio and Brunero Liseo</i>	
<b>Spherical regression with local rotations and implementation in R .....</b>	<b>269</b>
Regressione sferica con rotazioni locali ed implementazione in R	
<i>Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor</i>	
<b>A clustering method for network data to analyse association football playing styles .....</b>	<b>277</b>
Un metodo di raggruppamento per dati di rete finalizzato all'analisi degli schemi di gioco nel calcio	
<i>Jacopo Diqigiovanni</i>	
<b>Big data in longitudinal observational studies: how to deal with non-probability samples and technological changes.....</b>	<b>285</b>
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>	
<b>Smart Data For Smart Health.....</b>	<b>293</b>
Smart Data Per Smart Health	
<i>Clelia Di Serio, Ernst C. Wit, Elena Bottinelli and Roberto Buccione</i>	
<b>Detecting and classifying moments in basketball matches using sensor tracked data.....</b>	<b>297</b>
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>	
<b>Ordered response models for cyber risk .....</b>	<b>305</b>
Modelli a risposta ordinale per la valutazione del cyber risk	
<i>Silvia Facchinetti and Claudia Tarantola</i>	
<b>Functional data analysis-based sensitivity analysis of integrated assessment Models for climate change modelling .....</b>	<b>313</b>
Analisi di sensitività basata sull'analisi di dati funzionali per modelli di valutazione integrata dei cambiamenti climatici	
<i>Matteo Fontana, Massimo Tavoni and Simone Vantini</i>	
<b>Coupled Gaussian Processes for Functional Data Analysis.....</b>	<b>319</b>
Processi gaussiani per l'analisi dei dati funzionali	
<i>L. Fontanella, S. Fontanella, R. Ignaccolo, L. Ippoliti, P. Valentini</i>	
<b>Two-fold data streams dimensionality reduction approach via FDA .....</b>	<b>323</b>
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>	
<b>Statistical analysis of Sylt's coastal profiles using a spatiotemporal functional model .....</b>	<b>331</b>
<i>Rik Gijsman, Philipp Otto, Torsten Schlurmann, Jan Visscher</i>	
<b>Bootstrap prediction intervals for weighted TAR predictors .....</b>	<b>339</b>
Intervalli di previsione bootstrap per previsori ponderati per modelli TAR	
<i>Francesco Giordano and Marcella Niglio</i>	
<b>A rank graduation index to prioritise cyber risks .....</b>	<b>347</b>
Un indice di graduazione per assegnare livelli di priorità ai rischi informatici	
<i>Paolo Giudici and Emanuela Raffinetti</i>	
<b>Vector Error Correction models to measure connectedness of bitcoin exchange markets</b>	<b>355</b>
Modelli di Vector Error Correction per misurare la connessione delle piattaforme di scambio di bitcoin	
<i>Paolo Giudici and Paolo Pagnottoni</i>	
<b>Estimation of lineup efficiency effects in Basketball using play-by-play data.....</b>	<b>363</b>
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>	
<b>Trajectory clustering using adaptive squared distances .....</b>	<b>371</b>
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. Rodríguez-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'Ambr, D'Ambr 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 spatio-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 spatio-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>	



<b>A Prediction Method for Ordinal Consistent Partial Least Squares</b> .....	725
Un Metodo di Previsione per l'Algoritmo Ordinal Consistent Partial Least Squares	
<i>Gabriele Cantaluppi and Florian Schuberth</i>	
<b>Functional control charts for monitoring ship operating conditions and CO2 emissions based on scalar-on-function linear model</b> .....	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>	
<b>Predicting and improving smart mobility: a robust model-based approach to the BikeMi BSS</b> .....	737
Prevedere e migliorare la mobilità smart: un approccio robusto di classificazione applicato a BikeMi	
<i>Andrea Cappozzo, Francesca Greselin and Giancarlo Manzi</i>	
<b>Public support for an EU-wide social benefit scheme: evidence from Round 8 of the European Social Survey (ESS)</b> .....	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>	
<b>Revenue management strategies and Booking.com ghost rates: a statistical analysis</b> .....	751
Strategie di revenue management e Booking.com ghost rates: un'analisi statistica	
<i>Cinzia Carota, Consuelo R. Nava, Marco Alderighi</i>	
<b>Analysing international migration flows: a Bayesian network approach</b> .....	757
Analisi dei flussi migratori internazionali attraverso l'impiego di modelli grafici	
<i>Federico Castelletti and Emanuela Furfaro</i>	
<b>A sparse estimator for the function-on-function linear regression model</b> .....	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>	
<b>Robustness and fuzzy multidimensional poverty indicators: a simulation study</b> .....	769
Robustezza ed indicatori fuzzy multidimensionali della povertà: uno studio di simulazione	
<i>Michele Costa</i>	
<b>Text Based Pricing Modelling: an Application to the Fashion Industry</b> .....	775
Modellazione dei prezzi basata su dati testuali: un'applicazione all'industria fashion	
<i>Federico Crescenzi, Marzia Freo and Alessandra Luati</i>	
<b>Model based clustering in group life insurance via Bayesian nonparametric mixtures</b> .....	781
Raggruppamento basato sul modello nel settore assicurativo: un approccio bayesiano nonparametrico	
<i>Laura D'Angelo</i>	
<b>Smart Tools for Academic Submission Decisions: Waiting Times Modeling</b> .....	787
Strumenti "Smart" per sottoporre i manoscritti accademici: modelli per i tempi di attesa	
<i>Francesca De Battisti - Giancarlo Manzi</i>	
<b>On the Use of Control Variables in PLS-SEM</b> .....	793
Sull'Uso delle Variabili di Controllo nei PLS-SEM	
<i>Francesca De Battisti and Elena Siletti</i>	
<b>Partial dependence with copula and financial applications</b> .....	799
Dipendenza parziale con funzioni copula e applicazioni finanziarie	
<i>Giovanni De Luca, Marta Nai Ruscone and Giorgia Riveccio</i>	
<b>Exploring the relationship between fertility and well-being: What is smart?</b> .....	805
Esplorando la relazione tra fecondità e benessere: cosa c'è di smart?	
<i>Alessandra De Rose, Filomena Racioppi, Maria Rita Sebastiani</i>	
<b>Web-Based Data Collection and Quality Issues in Co-Authorship Network Analysis</b> .....	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>	
<b>A new regression model for bounded multivariate responses</b> .....	817
Un nuovo modello di regressione per risposte multivariate limitate	
<i>Agnese Maria Di Brisco, Roberto Ascarì, Sonia Migliorati and Andrea Ongaro</i>	
<b>Turning big data into smart data: two examples based on the analysis of the Mappa dei Rischi dei Comuni Italiani</b> .....	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>	

<b>Hidden Markov Model estimation via Particle Gibbs .....</b>	<b>829</b>
Stima di Hidden Markov Model tramite Particle Gibbs	
<i>Pierfrancesco Alaimo Di Loro, Enrico Ciminello and Luca Tardella</i>	
<b>A note on marginal effects in logistic regression with independent covariates .....</b>	<b>837</b>
Una nota sugli effetti marginali nella regressione logistica con covariate indipendenti	
<i>Marco Doretti</i>	
<b>DNA mixtures: a case study involving a Romani reference population .....</b>	<b>843</b>
Misure di DNA: un caso di studio riguardante una popolazione di riferimento dei Rom	
<i>Francesco Dotto, Julia Mortera and Vincenzo Pascali</i>	
<b>Pivotal seeding for K-means based on clustering ensembles .....</b>	<b>849</b>
Inizializzazione pivotale dell'algoritmo delle K-medie tramite raggruppamento con metodi di insieme	
<i>Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli</i>	
<b>Optimal scoring of partially ordered data, with an application to the ranking of smart cities</b>	<b>855</b>
Scoring ottimale di dati parzialmente ordinati, con un'applicazione al ranking delle smart city	
<i>Marco Fattore, Alberto Arcagni, Filomena Maggino</i>	
<b>Bounded Domain Density Estimation .....</b>	<b>861</b>
Stima della densità non-parametrica su domini bidimensionali limitati	
<i>Federico Ferraccioli, Laura M. Sangalli and Livio Finos</i>	
<b>Polarization and long-run mobility: yearly wages comparison in three southern European countries .....</b>	<b>867</b>
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>	
<b>Design of Experiments, aberration and Market Basket Analysis .....</b>	<b>873</b>
Pianificazione degli esperimenti, aberrazione e Market Basket Analysis	
<i>Roberto Fontana and Fabio Rapall</i>	
<b>Generalized Procrustes Analysis for Multilingual Studies .....</b>	<b>879</b>
Analisi Procrustiana Generalizzata per studi Multilingue	
<i>Alessia Forciniti, Michelangelo Misuraca, Germana Scepti, Maria Spano</i>	
<b>Prior specification in flexible models .....</b>	<b>885</b>
Specificazione delle prior in modelli flessibili	
<i>Maria Franco-Villoria, Massimo Ventrucci and Haavard Rue</i>	
<b>Modeling Cyclists' Itinerary Choices: Evidence from a Docking Station-Based Bike-Sharing System .....</b>	<b>889</b>
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>	
<b>A PARAFAC-ALS variant for fitting large data sets .....</b>	<b>895</b>
Una variante del PARAFAC-ALS per approssimare data set di grandi dimensioni	
<i>Michele Gallo, Violetta Simonacci and Massimo Guarino</i>	
<b>A Convex Mixture Model for Binomial Regression .....</b>	<b>901</b>
Un modello mistura convessa per la Regressione Binomiale	
<i>Luisa Galtarossa and Antonio Canale</i>	
<b>Blockchain as a universal tool for business improvement .....</b>	<b>907</b>
Blockchain come strumento universale per il miglioramento del business	
<i>Massimiliano Giacalone, Diego Carmine Sinitò, Emilio Massa, Federica Oddo, Enrico Medda, Vito Santarcangelo</i>	
<b>Seasonality in tourist flows: a decomposition of the change in seasonal concentration .....</b>	<b>913</b>
La stagionalità nei flussi turistici: una scomposizione della variazione nella concentrazione stagionale	
<i>Luigi Grossi and Mauro Mussini</i>	
<b>Are Real World Data the smart way of doing Health Analytics? .....</b>	<b>919</b>
Real World Data: la base di una nuova ricerca clinica?	
<i>Francesca Ieva</i>	
<b>Internet use and leisure activities: are all young people equal? .....</b>	<b>925</b>
Internet e tempo libero: i giovani sono uguali tra loro?	
<i>Giuseppe Lamberti, Jordi Lopez Sintas and Pilar Lopez Belbeze</i>	
<b>On a Family of Transformed Stochastic Orders .....</b>	<b>931</b>
Su una famiglia di ordinamenti stocastici trasformati	
<i>Tommaso Lando and Lucio Bertoli-Barsotti</i>	

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

<b>Betting on football: a model to predict match outcomes</b> .....	1027
<i>Scommettere sul calcio: un nuovo modello per prevedere l'esito delle partite</i>	
<i>Marco Petretta, Lorenzo Schiavon and Jacopo Diquigiovanni</i>	
<b>Estimation of dynamic quantile models via the MM algorithm</b> .....	1033
<i>Stima di modelli Quantilici Dinamici con algoritmo MM</i>	
<i>Fabrizio Poggioni, Mauro Bernardi, Lea Petrella</i>	
<b>The decomposition by subpopulations of the Pietra index: an application to the professional football teams in Italy</b> .....	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>	
<b>An Object Oriented Data Analysis of Tweets: the Case of Queen Elizabeth Olympic Park</b> .	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>	
<b>Bias reduced estimation of a fixed effects model for Expected Goals in association football</b> .....	1051
<i>Stima non distorta di un modello Expected Goal con effetti fissi nel calcio</i>	
<i>Lorenzo Schiavon and Nicola Sartori</i>	
<b>Looking for Efficient Methods to Collect and Geolocalise Tweets</b> .....	1057
<i>Alla ricerca di metodi efficienti per raccogliere e geolocalizzare tweet</i>	
<i>Stephan Schlosser, Daniele Toninelli and Silvia Fabris</i>	
<b>Principal ranking profiles</b> .....	1063
<i>Principal ranking profiles</i>	
<i>Mariangela Sciandra, Antonella Plaia</i>	
<b>A statistical model for voting probabilities</b> .....	1069
<i>Un modello statistico per le probabilità di voto</i>	
<i>Rosaria Simone, Stefania Capecchi</i>	
<b>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</b> .....	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>	
<b>Dealing with uncertainty in automated test assembly problems</b> .....	1083
<i>La gestione dell'incertezza nei problemi di assemblaggio automatizzato dei test</i>	
<i>Giada Spaccapanico Proietti, Mariagiulia Matteucci and Stefania Mignani</i>	
<b>Joint Models: a smart way to include functional data in healthcare analytics</b> .....	1089
<i>Modelli congiunti: un metodo per includere i dati funzionali nelle analisi in ambito sanitario</i>	
<i>Marta Spreafico, Francesca Ieva</i>	
<b>Bayesian multiscale mixture of Gaussian kernels for density estimation</b> .....	1095
<i>Stima di densità tramite misture bayesiane multiscala di kernel gaussiani</i>	
<i>Marco Stefanucci and Antonio Canale</i>	
<b>Dynamic Bayesian clustering of running activities</b> .....	1101
<i>Clustering Bayesiano dinamico di attività di corsa</i>	
<i>Mattia Stival and Mauro Bernardi</i>	
<b>Employment and fertility in couples: whose employment uncertainty matter most?</b> .....	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>	
<b>A Functional Data Analysis Approach to Study a Bike Sharing Mobility Network in the City of Milan</b> .....	1113
<i>Agostino Torti, Alessia Pini and Simone Vantini</i>	
<b>Multiresolution Topological Data Analysis for Robust Activity Tracking</b> .....	1119
<i>Giovanni Trappolini, Tullia Padellini, and Pierpaolo Brutti</i>	
<b>Semilinear regression trees</b> .....	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 <a href="mailto:ivan.sciascia@unito.it">ivan.sciascia@unito.it</a>, Gennaro Carotenuto <a href="mailto:gennaro.carotenuto@unito.it">gennaro.carotenuto@unito.it</a>, Andrea Genre <a href="mailto:andrea.genre@unito.it">andrea.genre@unito.it</a>, Università di Torino Dipartimento di Scienze della vita e biologia dei sistemi, viale Mattioli 25, 10125 Torino</i>	

# Density-based Algorithm and Network Analysis for GPS Data

## *Algoritmi di Cluster e Reti per lo studio di dati GPS*

Antonino Abbruzzo, Mauro Ferrante, Stefano De Cantis

**Abstract** The use of advanced global positional system (GPS) trackers has emerged as a novel technology in data collection of units movements. GPS data contain a large amount of information since the signals of the units are recorded almost in real time. The analysis of GPS data can be carried on several aspects of the spatial movements. In this study, we focus on statistical methods for the identification of points of interests and the analysis of the network of movements for GPS data. In particular, a density cluster-based algorithm is applied to summarize the vast amount of information and to find the most relevant points of attractions. A directed network synthesizes the individual unit path by using the latter information. Finally, we aggregate the unit paths in a weighted directed network which is studied through network analysis. We apply the proposed approach to a case study on cruise passengers' movements in an urban context.

**Abstract** *La diffusione dei sistemi di localizzazione GPS offre numerose opportunità per la raccolta di dati di movimento. I dati GPS presentano diversi elementi di complessità derivanti anche dall'elevato dettaglio temporale e territoriale. Numerosi sono gli aspetti che possono essere presi in esame per tale tipologia di dati. Il presente studio propone un approccio statistico basato sull'identificazione dei punti di attrazione e sullo studio dei network. In particolare, viene proposto un algoritmo di identificazione di cluster, sulla base della densità di punti, che vengono sintetizzati in un network che riassume il comportamento individuale. In un secondo step, i movimenti complessivi sono aggregati ed analizzati tramite la network analysis. L'approccio proposto è applicato allo studio dei movimenti di croceristi in contesti urbani.*

**Key words:** Spatial-Temporal Data, Cluster-Based Method, Tourists' Behaviors, Destination Management.

---

Department of Economics, Business and Statistics  
University of Palermo, Viale delle Scienze, building 13, 90128 - Palermo e-mail: antonino.abbruzzo@unipa.it; stefano.decantis@unipa.it

Department of Culture and Society  
University of Palermo, Viale delle Scienze, building 15, 90128 - Palermo e-mail: mauro.ferrante@unipa.it

## 1 Introduction

Nowadays, GPS devices have become of small size, not bulky, equipped with significant autonomy and, what matters most, once activated manage to memorize the geographical coordinates in which the statistical unit is at a given moment. Once the experience is over, it will be possible to download the data and to analyze the route taken by the unit. These devices have several advantages and give the opportunity to collect high-quality data, that are very accurate both in terms of temporal (seconds) and spatial (meters) resolution. Moreover, this type of data provides information on the unit's movement not influenced by units' perceptions or other issues (e.g., recall bias) which generally affect traditional survey instruments (e.g., diaries or questionnaires).

The analysis of GPS data can be conducted on several aspects of the spatial movements. In this paper, we focus our attention on statistical methods for the identification of points of interests (POIs) and the units' networks of movements. Moreover, we aggregate these units' movements networks to produce a directed weighted network which can be studied through network analysis. Various techniques have been proposed to detect POIs and visualize patterns, including density estimation grid-based aggregation, spatial movement sequence, network analysis [4] and algorithms for clustering spatial data [5].

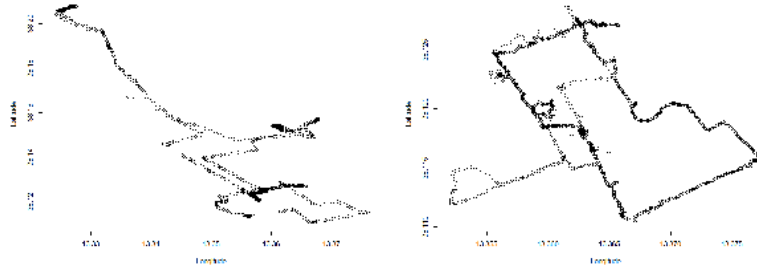
We introduce a statistical approach based on density based-cluster algorithm (DBSCAN [2]) and networks analysis. The clustered approach on the GPS data recovers the "points of interests. The network analysis summarizes units density cluster-based data to access to the most relevant characteristics of the units movements. Finally, we apply the proposed approach to the case of cruise passengers' mobility in the city of Palermo.

## 2 Statistical Methods

In this section, we first describe the DBSCAN algorithm which summarizes a large amount of information and find the most relevant points of interests. Secondly, we define the directed network approach to summarize units density cluster-based data to access to the most significant characteristics of the units movements. Finally, we derive a weighted direct graph which summarizes the behaviors of all the sampled units.

### 2.1 Density Cluster-Based Model for GPS data

Let  $D_j^{(i)} = (t_j^{(i)}, long_j^{(i)}, lat_j^{(i)})$ , where  $i = 1, \dots, N$ ,  $j = 1, \dots, n_i$ ,  $t_{j+1}^{(i)} - t_j^{(i)} = c$  and  $c$  is a constant, be the temporal-spatial movement of the  $i$ -th unit. Figure 1 shows two temporal-spatial sequences referred to two different units. In this Figure, we may recognize the points of interests for each of the units, by looking at the higher points density in some locations of the path. DBSCAN is a density-based algorithm designed to discover arbitrary-shaped clusters in any database  $D_j^{(i)}$  and at the same



**Fig. 1** Two illustrative units' behaviors.

time to distinguish noise points. These clusters are called “points of interests” and they can be interpreted as a set of places a unit visited for a certain amount of time.

Let us introduce some concepts before describing the DBSCAN algorithm. The  $\varepsilon$ -neighbourhood of a point  $p$  is defined by  $ne_\varepsilon(p) = \{q \in D_{ij} : dist(p, q) \leq \varepsilon\}$ , where  $dist(p, q)$  is a distance function (e.g., Manhattan Distance, Euclidean Distance). If the cardinality of an  $\varepsilon$ -neighbourhood of a point  $p$ , i.e.  $|ne_\varepsilon(p)|$ , is at least greater than a minimum number ( $MinPts$ ) then  $p$  is a **core point**. A point  $p$  is **directly density-reachable** from the object  $q$  with respect to  $\varepsilon$  and  $MinPts$  if  $p \in ne_\varepsilon(q)$  and  $|ne_\varepsilon(q)| \geq MinPts$ . A point  $p$  is **density-reachable** from the object  $q$  with respect to  $\varepsilon$  and  $MinPts$  if there is a chain  $p_1, \dots, p_l, p_1 = q, p_l = p$  such that  $p_{l+1}$  is directly density-reachable from  $p_l$ . An object  $p$  is **density-connected** to object  $q$  with respect to  $\varepsilon$  and  $MinPts$  if there is an object  $o$  such that both,  $p$  and  $q$  are density reachable from  $o$  with respect to  $\varepsilon$  and  $MinPts$ . A **cluster**  $C$  is a non-empty subset of  $D_j^{(i)}$  satisfying the following maximality and connectivity requirements:

- $\forall p, q$ : if  $q \in C$  and  $p$  is density-reachable from  $q$  with respect to  $\varepsilon$  and  $MinPts$ , then  $p \in C$ ;
- $\forall p, q \in C$ :  $p$  is density-connected to  $q$  with respect to  $\varepsilon$  and  $MinPts$ .

An object  $p$  is a **noise** object if it is not a core object but density-reachable from another core object.

The algorithm starts with the first point  $p$  in the database  $D_j^{(i)}$ , and it retrieves all the neighbors of a point  $p$  with respect to  $\varepsilon$  and  $MinPts$ . If  $p$  is a core point, this procedure yields a cluster concerning  $\varepsilon$  and  $MinPts$ . If  $p$  is a border point, no points are density-reachable from  $p$ , and the DBSCAN algorithm proceeds in considering the next point of the database.

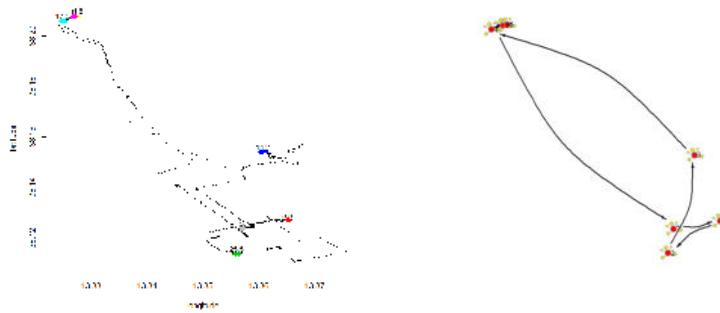
## 2.2 Network of a DBSCAN object

The application of the DBSCAN algorithm defined in Section 2.1 produces, for each  $\mathbf{D}^{(i)} = \{D_1^{(i)}, \dots, D_{n_i}^{(i)}\}$ , a DBSCAN object, i.e. a set of clusters  $C^{(i)} = \{C_1^{(i)}, \dots, C_{k_i}^{(i)}\}$  corresponding to a set of spatial coordinates of unit  $i$  which satisfy the maximality



and connectivity requirements. For each  $C_m^{(i)}$ ,  $m = 1 \dots k_i$ , we can associate the time spent by the unit  $i$  into the cluster  $m$  by multiplying the cardinality of the cluster by the constant  $c$ .

A directed graph for a unit  $i$  is defined as  $G^{(i)} = (V^{(i)}, E^{(i)})$  where  $V^{(i)}$  is a set of nodes which corresponds to the set of clusters or points of interests  $\bar{C}^{(i)} = \{1, \dots, k_i\}$ ,  $E^{(i)}$  is a set of links (connections between points of interest for the  $i$ -th unit), where  $e_{kl}^{(i)} = 1$  if the unit goes from node  $k$  to node  $l$  and 0 otherwise. The latter information is recovered thanks to the temporal ordering of the spatial movements.



**Fig. 2** Raw GPS data and clusters obtained with the DBSCAN algorithm (on the left); the numbers over each cluster represent the times spent by the unit into that cluster. Nodes and links (on the right) represent points of interest and sequence of visit, respectively, for the considered unit.

Figure 2 shows an example of the application of DBSCAN to  $\mathbf{D}_1$ , i.e. the temporal-spatial data for unit 1 (left part of Figure 2), and the recovered directed network (right part of Figure 2). In this example DBSCAN estimates seven clusters  $C^{(1)} = \{C_1^{(1)}, \dots, C_7^{(1)}\}$  with times spent in each cluster that are equal to:  $t_1^{(1)} = 5.5, t_2^{(1)} = 25.5, t_3^{(1)} = 14.2, t_4^{(1)} = 17.5, t_5^{(1)} = 9, t_6^{(1)} = 4.8, t_7^{(1)} = 5$  minutes, respectively. The directed network has seven vertices  $V = \{1, \dots, 7\}$  and the directed links show the path of the  $i$ -th unit. In this example, we can recover from the temporal information that unit 1 starts from node 1 and chooses the following itinerary 1 - 2 - 3 - 4 - 5 - 6 - 5 - 4 - 7 - 1. Note that the positions of each node in the network correspond to the centroid of each cluster. According to this approach, the information about the noise points (i.e., the streets the units followed during their path) is lost, being not of interest for the present study.

### 2.3 Network Analysis

Let  $N$  be the number of collected sample units. So far, we have described an approach to derive a set of clusters and directed networks from the temporal-spatial GPS data  $\mathbf{D}^{(i)}$  for each unit  $i$ . The application of the DBSCAN to  $\mathbf{D}^{(i)}$ ,  $i = 1, \dots, N$

produces the spatial data  $\mathbf{C} = \{C^{(1)}, \dots, C^{(N)}\}$  which allows to reduce the number of points collected for the sample units by discarding the noise points.

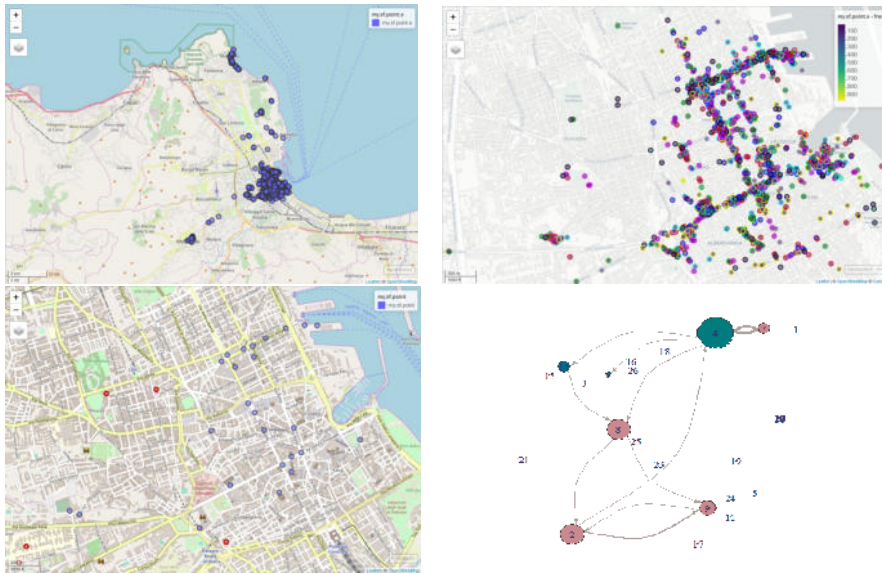
In this section, we define a weighted directed graph which summarizes the behaviors of all the sampled units. The weighted directed graph is a triplet  $G = (V, E, W)$ , where the set  $V$  represents the POIs of the collected sample units obtained by applying the DBSCAN algorithm to  $\mathbf{C}$ . In other words, the DBSCAN algorithm is applied at an aggregate level. Then the set of edges  $E$  represents the transitions from a POI to another one. An edge  $e_{lk} = 1$  if a unit goes from cluster  $l$  to cluster  $k$  and zero otherwise. Finally, the weight matrix  $W$  represents the number of units that go from a node to another one. Specifically,  $w_{lk}$  indicates the number of units that went from the node  $l$  to the node  $k$  and the diagonal of  $W_{ll}$  represents the number of units that spent a certain amount of time into the cluster  $l$ .

The identification of relevant nodes of the network  $G$  is one of the main applications of network analysis. We consider two measures to establish nodes importance: **degree centrality** which count of the number of directed connections of a node, and the **page rank centrality**, a variant form of eigenvector centrality, which has been used to evaluate the influence power of nodes on surrounding nodes [4].

### 3 Empirical Application

In this section, we apply the proposed methods on a sample of 303 GPS tracks related to cruiser passengers disembarked in the city of Palermo in 2014. Details on data collection procedures and other characteristics of the survey may be found in [1].

The implementation of the density-based algorithm described in Section 2.1 at the individual level allowed for a synthesis of the main POIs visited by each cruise passengers. Figure 3 shows, on the upper left side, the clusters of the entire city of Palermo which can be classified in three zones: Mondello (a beach place), Monreale (a mountain place) and the city center (historical area). The upper right side of the Figure shows, by zooming at the city center, many POIs. In order to recover the most important attraction points, we applied again the DBSCAN algorithm but at an aggregate level, i.e. on the clusters of Figure 3 upper right side. We obtained the POIs showed in the left lower part of Figure 3. In the lower right part of Figure 3, we show the weighted directed networks of the main POIs visited by the units of our sample. Note that we cut the edge if  $w_{lk} \leq 20$ , which means that an edge will be present if more than 20 units go from node  $l$  to node  $k$ . The size of each POI is proportional to its degree. The colors of the nodes are associated with their measure of page rank centrality. The Figure shows a path that starts from the harbor (node 4) and ends to the Palermo's Real palace (node 8) which is one of the main attraction in the city. Node 8, which represents the Palermo' cathedral, looks like another relevant POI.



**Fig. 3** Centroid Cluster recovered by using Algorithm 1. On the left side the cluster of the entire city of Palermo and on the right side the zoom on the city center.

## 4 Conclusion

In this paper, we proposed a density-based algorithm to reduce the complexity of GPS spatial-temporal data of a unit by finding the clusters which we interpreted as points of interest. These POIs can be summarized in a network. The information loss is minimal since the network gives the entire path of the unit unless a specific interest on the route followed by the unit is of importance. Moreover, we used the network analysis to merge the different units behaviors. At the individual level, the proposed methodology provides a synthesis of unit behavior in terms of main POIs visited and of the sequence of visits. At the aggregate level, it provides a structure of the main POIs, and on the general chain of visits, as well as on the strength of the links between each POIs in terms of flows of visits. The proposed methodology does not require any prior knowledge for the identification of the POIs, being easily replicable in different contexts.

## References

1. De Cantis, S., Ferrante, M., Kahani, A., & Shoval, N. (2016). Cruise passengers' behavior at the destination: Investigation using GPS technology. *Tourism Management*, 52, 133-150.
2. Ester, M., Kriegel, H.P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, 96(33), 226-231.
3. Sia-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science*, 30(5), 881-906.
4. Sugimoto, K., Ota, K., & Suzuki, S. (2019). Visitor Mobility and Spatial Structure in a Local Urban Tourism Destination: GPS Tracking and Network analysis. *Sustainability*, 11(3), 919.
5. Varghese, M., Unnikrishnan, A., & Jacob, K. (2013). Spatial clustering algorithms—An overview. *Asian Journal of Computer Science And Information Technology*, 3(11), 1-8.