

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.	141
Tecnologie e data science per una salute migliore sia a livello individuale che di popolazione.	
<i>Stefano Campostrini and Lucia Zanotto</i>	
Temporal sentiment analysis with distributed lag models	149
Analisi temporale del "sentiment" con modelli a lag distribuiti	
<i>Carrannante M., Mattera R., Misuraca M., Scepi G., Spano M.</i>	
A statistical investigation on the relationships among financial disclosure, sociodemographic variables, financial literacy and retail investors' risk assessment ability	157
Indagine empirica sulle relazioni tra prospetti per la diffusione di informazioni finanziarie, variabili sociodemografiche, educazione finanziaria e abilità di valutazione del rischio	
<i>Rosella Castellano, Marco Mancinelli and Pasquale Sarnacchiaro</i>	
Bayesian Model Comparison based on Wasserstein Distances.....	167
Confronto di Modelli Bayesiani tramite Distanze di Wasserstein	
<i>Marta Catalano, Antonio Lijoi and Igor Prünster</i>	
Hierarchical Clustering and Dimensionality Reduction for Big Data	173
Clustering e Riduzione Dimensionale Gerarchici per Dati di Grandi Dimensioni	
<i>Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria</i>	
ICOs success drivers: a textual and statistical analysis.....	181
Fattori di successo nelle ICOs: un'analisi testuale e statistica	
<i>Paola Cerchiello and Anca Mirela Toma</i>	
Small area estimators with linked data.....	189
Stimatori per piccole aree nel caso di dati ottenuti attraverso il record linkage	
<i>Chambers Raymond and Fabrizi Enrico and Salvati Nicola</i>	
Optimal Portfolio Selection via network theory in banking and insurance sector	197
<i>Gian Paolo Clemente, Rosanna Grassi and Asmerilda Hitaj</i>	
Matching error(s) and quality of statistical matching in complex surveys.....	205
Errori di matching e qualità del matching statistico in indagini complesse	
<i>Pier Luigi Conti and Daniela Marella</i>	
Hotel search engine architecture based on online reviews' content.....	213
Un motore di ricerca per gli hotel basato sulle recensioni online	
<i>Claudio Conversano, Maurizio Romano and Francesco Mola</i>	
Economic Crisis and Earnings Management: a Statistical Analysis	219
Crisi Economica e Gestione degli Utili: un'Analisi Statistica	
<i>C. Cusatelli, A.M. D'Uggento, M. Giacalone, F. Grimaldi</i>	
A Comparison of Nonparametric Bivariate Survival Functions.....	227
Confronto tra stimatori non-parametrici della funzione di sopravvivenza bivariata	
<i>Hongsheng Dai and 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 Diquigiovanni</i>	
Big data in longitudinal observational studies: how to deal with non-probability samples and technological changes.....	285
I Big data negli studi longitudinali: come trattare campioni non probabilistici e cambi di tecnologia	
<i>Clelia Di Serio, Luca Del Core, Eugenio Montini and Andrea Calabria</i>	
Smart Data For Smart Health.....	293
Smart Data Per Smart Health	
<i>Clelia Di Serio, Ernst C. Wit, Elena Bottinelli and Roberto Buccione</i>	
Detecting and classifying moments in basketball matches using sensor tracked data.....	297
Una procedura per identificare e classificare momenti di gioco in pallacanestro con l'uso di dati sensori.	
<i>Tullio Facchinetti and Rodolfo Metulini and Paola Zuccolotto</i>	
Ordered response models for cyber risk	305
Modelli a risposta ordinale per la valutazione del cyber risk	
<i>Silvia Facchinetti and Claudia Tarantola</i>	
Functional data analysis-based sensitivity analysis of integrated assessment Models for climate change modelling	313
Analisi di sensibilità basata sull'analisi di dati funzionali per modelli di valutazione integrata dei cambiamenti climatici	
<i>Matteo Fontana, Massimo Tavoni and Simone Vantini</i>	
Coupled Gaussian Processes for Functional Data Analysis.....	319
Processi gaussiani per l'analisi dei dati funzionali	
<i>L. Fontanella, S. Fontanella, R. Ignaccolo, L. Ippoliti, P. Valentini</i>	
Two-fold data streams dimensionality reduction approach via FDA	323
Un approccio a due fasi per la riduzione di dimensionalità di data streams via FDA	
<i>F. Fortuna, T. Di Battista and S.A. Gattone</i>	
Statistical analysis of Sylt's coastal profiles using a spatiotemporal functional model	331
<i>Rik Gijssman, Philipp Otto, Torsten Schlurmann, Jan Visscher</i>	
Bootstrap prediction intervals for weighted TAR predictors	339
Intervalli di previsione bootstrap per previsori ponderati per modelli TAR	
<i>Francesco Giordano and Marcella Niglio</i>	
A rank graduation index to prioritise cyber risks.....	347
Un indice di graduazione per assegnare livelli di priorità ai rischi informatici	
<i>Paolo Giudici and Emanuela Raffinetti</i>	
Vector Error Correction models to measure connectedness of bitcoin exchange markets	355
Modelli di Vector Error Correction per misurare la connessione delle piattaforme di scambio di bitcoin	
<i>Paolo Giudici and Paolo Pagnottoni</i>	
Estimation of lineup efficiency effects in Basketball using play-by-play data.....	363
L'uso dei dati del play-by-play per la stima degli effetti di quintetto nella pallacanestro	
<i>Luca Grassetti, Ruggero Bellio, Giovanni Fonseca and Paolo Vidoni</i>	
Trajectory clustering using adaptive squared distances	371
Clustering di traiettorie attraverso distanze adattative quadratiche	
<i>Antonio Irpino</i>	

Bayesian Analysis of Privacy Attacks on GPS Trajectories	379
<i>Analisi Bayesiana degli Attacchi alla Privacy su Traiettorie GPS</i>	
<i>Sirio Legramanti</i>	
Data Analytics in the Insurance Industry: Market trends and lessons from a use case customer predictive modelling	387
<i>Data Analytics nel settore assicurativo: principali trend e considerazioni da un caso d'uso applicato alla predizione del comportamento degli assicurati</i>	
<i>Cristian Losito and Francesco Pantisano</i>	
BasketballAnalyzeR: the R package for basketball analytics	395
<i>BasketballAnalyzeR: il pacchetto R per l'analisi dei dati nella pallacanestro</i>	
<i>Marica Manisera, Marco Sandri and Paola Zuccolotto</i>	
Data Integration by Graphical Models	403
<i>Utilizzo dei modelli grafici per l'integrazione dei dati</i>	
<i>Daniela Marella and Paola Vicard and Vincenzina Vitale</i>	
A two-part finite mixture quantile regression model for semi-continuous longitudinal data	409
<i>Maruotti Antonello, Merlo Luca and Petrella Lea</i>	
Multivariate change-point analysis for climate time series	415
<i>Analisi di change-point 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 Zurlò 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-esimo vicino più vicino	
<i>Siirio 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'Ambra, D'Ambra Antonello</i>	
A statistical space-time functional model for air quality analysis and mapping.....	593
Un modello statistico spazio-tempo funzionale per l'analisi e la mappatura della qualità dell'aria	
<i>Yaqiong Wang, Alessandro Fassò and Francesco Finazzi</i>	
Tempering and computational efficiency of Bayesian variable selection.....	599
Tempering e l'efficienza computazionale della selezione bayesiana delle variabili	
<i>Giacomo Zanella and Gareth O. Roberts</i>	
Dimensions and links for Hate Speech in the social media	607
Dimensioni e legami per i discorsi di odio nei social media	
<i>Emma Zavarrone, Guido Ferilli</i>	

Section 3. Contributed Papers

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

Application and validation of dynamic Poisson models to measure credit contagion	629
Applicazione e validazione di modelli di Poisson dinamici per misurare il contagio nel credito <i>Arianna Agosto and Emanuela Raffinetti</i>	
Monitoring SDGs at territorial level: the case of Lombardy.....	637
Il monitoraggio degli SDGs a livello territoriale: il caso della Lombardia <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
Il Metodo degli Esperti per la previsione di serie temporali multivariate e periodiche, di dimensione elevata <i>Giacomo Aletti, Marco Bellan and Alessandra Micheletti</i>	
Regression with time-dependent PDE regularization for the analysis of spatio-temporal data	649
Regressione con regolarizzazione di PDE tempo dipendenti per modellizzare dati spazio-temporali <i>Eleonora Annone, Laura Azzimonti, Fabio Nobile, Laura M. Sangalli</i>	
A network analysis of museum preferences: the Firenzecard experience.....	653
Un'analisi di rete delle preferenze museali: l'esperienza della Firenzecard <i>Silvia Bacci, Bruno Bertaccini, Roberto Dinelli, Antonio Giusti, and Alessandra Petrucci</i>	
A statistical learning approach to group response categories in questionnaires	659
Un approccio basato sull'apprendimento statistico per raggruppare le categorie di risposta nei questionari <i>Michela Battauz</i>	
Tree-based Functional Data Analysis for Classification and Regression.....	665
Alberi di Classificazione e Regressione per dati Funzionali <i>Edoardo Belli, Enrico Ragaini, Simone Vantini</i>	
PDE-regularized regression for anisotropic	669
spatial fields Regressione con regolarizzazione differenziale per campi spaziali anisotropi <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
Statistica classica nell'era dei big data. Verosimiglianza e modelli non regolari <i>Alessandra R. Brazzale and Valentina Mameli</i>	
Bayesian Variable Selection for High Dimensional Logistic Regression	685
Selezione bayesiana delle variabili nel modello di regressione logistica ad alta dimensionalità <i>Claudio Busatto, Andrea Sottosanti and Mauro Bernardi</i>	
Bayesian modeling for large spatio-temporal data: an application to mobile networks	691
Modelli bayesiani per grandi dataset spazio-temporali: un'applicazione a dati di telefonia mobile <i>Annalisa Cadonna, Andrea Cremaschi, Alessandra Guglielmi</i>	
A Mathematical Framework for Population of Networks: Comparing Public Transport of Different Cities.	697
Un approccio matematico all'analisi di una popolazione di networks: come confrontare il sistema di trasporto pubblico di diverse città. <i>Anna Calissano, Aasa Feragen, Simone Vantini</i>	
How Important Discrimination is for the Job Satisfaction of Immigrants in Italy: A Counterfactual Approach	703
Quanto influisce la discriminazione sulla soddisfazione lavorativa degli immigrati in Italia: un approccio controfattuale <i>Maria Gabriella Campolo, Antonino Di Pino and Michele Limosani</i>	
Unfolding the SEcrets of LongEvity: Current Trends and future prospects (SELECT)	709
A path through morbidity, disability and mortality in Italy and Europe <i>Stefano Campostrini, Daniele Durante, Fabrizio Faggiano and Stefano Mazzucco</i>	
Galaxy color distribution estimation via dependent nonparametric mixtures	713
Stima della distribuzione del colore delle galassie via misture nonparametriche dipendenti <i>Antonio Canale, Riccardo Corradin and Bernardo Nipoti</i>	
A case for order optimal matching: a salary gap study	719
Un algoritmo di matching ottimale ordinato per un studio sulle differenze salariali <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 Fuccella, Susanna Zaccarin</i>	
A new regression model for bounded multivariate responses	817
Un nuovo modello di regressione per risposte multivariate limitate	
<i>Agnese Maria Di Brisco, Roberto Ascari, Sonia Migliorati and Andrea Ongaro</i>	
Turning big data into smart data: two examples based on the analysis of the Mappa dei Rischi dei Comuni Italiani	823
Trasformare i big data in smart data: due esempi di analisi della Mappa dei Rischi dei Comuni Italiani	
<i>Oleksandr Didkovskiy, Alessandra Menafoglio, Piercesare Secchi, Giovanni Azzone</i>	

Hidden Markov Model estimation via Particle Gibbs	829
Stima di Hidden Markov Model tramite Particle Gibbs	
<i>Pierfrancesco Alaimo Di Loro, Enrico Ciminello and Luca Tardella</i>	
A note on marginal effects in logistic regression with independent covariates	837
Una nota sugli effetti marginali nella regressione logistica con covariate indipendenti	
<i>Marco Doretti</i>	
DNA mixtures: a case study involving a Romani reference population	843
Mixture 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
Massimiliano Giacalone, Diego Carmine Sinitò, Emilio Massa, Federica Oddo, Enrico Medda, Vito Santarcangelo	
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
Ricerca stocastica Bayesiana per modelli grafici a catena Ising	
<i>Andrea Lazzarini · Monia Lupporelli · Francesco C. Stingo</i>	
On the statistical design of parameters for variables sampling plans based on process capability index Cpk	941
Progettazione statistica dei parametri per il piano di campionamento per variabili basate sull'indice di capacità di processo Cpk	
<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
Nowcasting degli arrivi turistici stranieri usando Google Trends: un'applicazione nella città di Firenze, Italia	
<i>Alessandro Magrini</i>	
Inclusive growth in European countries: a cointegration analysis	953
La crescita inclusiva nei paesi europei: un'analisi di cointegrazione	
<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
ESCO il linguaggio europeo del lavoro: uno strumento concettuale ed operativo per le politiche del lavoro in contesti complessi	
<i>Cristina Martelli, Laura Grassini, Adham Kahlawi, Maria Flora Salvatori, Lucia Buzzigoli</i>	
Hidden Markov Models for High Dimensional Data	965
Hidden Markov Models per dati ad alta dimensionalità	
<i>Martino, A., Guatteri, G., Paganoni, A.M.</i>	
Classification of Italian classes via bivariate semi parametric multilevel models	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>	
Data Mining Application to Healthcare Fraud Detection: Two-Step Unsupervised Clustering Method for Outlier Detection with Administrative Databases.....	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>	
Multivariate analysis and biodiversity partitioning of a demersal fish community: an application to Lazio coast	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>	
Latent Markov models with discrete separate cluster random effects on initial and transition probabilities.....	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>	
Unsuitability of likelihood-based asymptotic confidence intervals for Response-Adaptive designs in normal homoscedastic trials	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 Zagoraiou</i>	
Local Hypothesis Testing for Functional Data: Extending False Discovery Rate to the Functional Framework.....	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>	
Educational mismatch and attitudes towards migration in Europe.....	1009
Disallineamento fra formazione e lavoro e atteggiamenti verso le migrazioni in Europa	
<i>Marco Guido Palladino and Emiliano Sironi</i>	
Soft thresholding Bayesian variable selection for compositional data analysis.....	1015
Selezione di Variabili Bayesiana con funzioni di soglia per l'analisi di dati di composizione	
<i>Matteo Pedone, Francesco C. Stingo</i>	
Sentiment-driven investment strategies: a practical example of AI-powered engines in a corporate setting	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>	

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

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

Section 4. Posters

Modelling Hedonic Price using semiparametric M-quantile regression	1147
Regressione m-quantilica semiparametrica per la modellizzazione dei prezzi edonici	
<i>Riccardo Borgoni, Antonella Carcagni, Alessandra Michelangeli, Nicola Salvati</i>	
Bayesian mixed latent factor model for multi-response marine litter data with multi-source auxiliary information	1153
Modello bayesiano misto a fattori latenti per l'abbondanza di rifiuti marini con informazioni ausiliarie di diversa provenienza	
<i>Crescenza 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 Universit 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 Universit Cattolica del Sacro Cuore, Universit degli Studi di Milano Bicocca, Universit Bocconi, Universit “Vita e salute” San Raffaele, Politecnico di Milano and Universit 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 *Fuoriconvegno* 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 *fuoriconvegno* 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>



Section 1. Plenary Sessions and Round Table

Shallow Learning for Data Science

Antonio Canale

Abstract One of the most recent breakthrough in data analysis is represented by the so called deep learning methods. Despite the huge hype surrounding deep learning, there is nowadays a strong need for statistical models that are “shallow”, i.e. models that are able to balance the modern quest for flexibility with parsimony and interpretability and that are able to account for uncertainty quantification. In this short paper I will argue that despite the surprising performance of deep learning in a variety of situations, this is not the ultimate general-purpose tool for data analysis. To argue this, I will focus on three applications where both the specific questions and the data structures requested a “smart” model specification tailored for each specific situation. Despite different in terms of goals and type of data, all these models share the common feature of providing useful and interpretable insights on the problems at hand.

Abstract *Uno dei più recenti e innovativi approcci all'analisi dei dati è rappresentato dai cosiddetti metodi di deep learning. Nonostante l'enorme clamore che circonda questo approccio, c'è ancora estremo bisogno di modelli statistici che non siano necessariamente “deep”, ovvero di modelli capaci di bilanciare l'attuale richiesta di flessibilità con caratteristiche come parsimonia e interpretabilità e che siano in grado di quantificare correttamente l'incertezza delle stime. In questo breve articolo viene discusso che, nonostante le sorprendenti prestazioni dei modelli di deep learning in svariati contesti, questi non rappresentano lo strumento definitivo per qualsivoglia analisi di dati. Per argomentare questo, vengono presentate tre applicazioni dove sia le domande specifiche che le strutture dati hanno richiesto lo sviluppo e l'applicazione di specifici modelli statistici. Sebbene differenti negli obiettivi e nelle strutture dati, questi modelli hanno in comune la capacità di fornire chiare informazioni relative ai problemi in questione.*

Key words: Deep learning, density regression, quantitative risk assessment, extreme value analysis, functional data analysis, functional regression

Antonio Canale

Dipartimento di Scienze Statistiche, Università degli Studi di Padova , e-mail: canale@stat.unipd.it

1 Introduction

One of the most recent breakthrough and successful tool for data analysis is represented by the so called deep learning methods [18, 30]. Deep learning is essentially a statistical model that builds over the concept of neural network, a formulation that first appeared at the end of the 50's [28] but that only recently became extremely popular thanks to the increase in computational power—i.e. hardware—and data size. In the last few years, not only deep learning proved to be competitive in a variety of tasks and applications, but it also proved to be quite successful in terms of marketing with abundant press exposure [23] and related interest also in the general public. All these reasons boosted deep learning that quickly became the go-to solution for many companies and for some scientific fields as well.

There is a huge hype surrounding deep learning but it is important to understand that this set of techniques cannot be the ultimate general-purpose solution for any data analysis. Also in situations in which these methods are supposed to give their best performance, there are many funny examples on the actual implications of their implementation [see, e.g. 19]. These examples need to warn practitioners about the risk of using these methods blindly as, after all, the term *deep* in deep learning “refers to a technical, architectural property [...] rather than a conceptual one” [22]. The scientific literature also started to question when and why these methods fail even in settings where they normally proved to be extremely successful (an entire conference has been held on the theme of deep learning failures, see [17]).

Deep learning has terrific performance in classification and regression tasks where the final goal is prediction and not inference—i.e. understanding of the process that generated the data and accounting for the intrinsic uncertainty of estimation. For this reason it is clear that if the goal of a data analysis is to understand the relationships between some variables or, even more, under a causal inference approach, the causality effect of a variable on a specific output, these methods are inherently inappropriate.

Deep learning has become extremely popular nowadays—the so called *big data* era—thanks to the availability of huge data sets. Their predictive performance is directly related to the size of the training data. Not only the size of the data set is important in determining successful results but also the quality of the data with the training set being a representative sample of future observations to be predicted. It is clear that if the sample is small to moderate in size and/or there is some (ignored) selection bias, the results of any deep learning algorithm may be not very precise or strongly wrong.

There are many situations and applications, nowadays, in which the substantive question, the data type, or the data size make deep learning methods not adequate for the reasons discussed above. In these situations there is still a strong need for statistical models that—in contrast to the extremely complex and overparametrized deep learning models—are “shallow,” i.e. models that are able to balance the modern quest for flexibility with parsimony and interpretability and that are able to account for uncertainty quantification. In what follows I will give three examples related to three research projects that, to some extent, support this thesis.

2 Smart statistics for smart applications?

In this section I will discuss three recent applications related to quantitative risk assessment, extreme daily rainfall, and bidding in energy markets, respectively. In each case the specific question and/or the data structure requested a model specification tailored to the specific application. Despite different in terms of goals and data structures, all these models share the common feature of providing useful and interpretable insights on the problems at hand without requiring too strict model assumptions.

2.1 Balancing flexibility with parsimony: a convex mixture regression model for quantitative risk assessment

Consider the context of environmental applications in which it is of interest to model how the distribution of a health outcome changes with a predictor. As motivating application, I focus on the analysis of a data set obtained from a sub-study of the US Collaborative Perinatal Project [20], a large prospective study of US pregnant women and children. In this data set, DDE—a persistent metabolite of DDT— was measured in the maternal serum during the pregnancy and the goal of the analysis is to study how the distribution of the gestational age at delivery—henceforth y —varies with DDE—henceforth x . Here, as in many other similar studies, the general interest is in relating dose to the risk of an adverse health outcome, with such dose–response models forming the basis of quantitative risk assessment [24]. The analysis presented here is a summary of the paper by Canale, Durante, and Dunson [2].

To give some background, typical approaches in quantitative risk assessment, rely on parametric models. For example, a common assumption is $y \sim N(\mu(x), \sigma^2)$, where $\mu(x)$ is defined as

$$\mu(x) = \mu_0 + (\mu_\infty - \mu_0)\psi(x; \lambda)$$

with μ_0 and μ_∞ denoting the expectations of y at $x = 0$ and $x \rightarrow \infty$, respectively, whereas $\psi(x; \lambda)$ represents a monotone nondecreasing dose–response function. While being interpretable, it is evident that the Gaussian assumption is clearly violated in many situations, including our motivating application as can be noticed in Figure 1. Alternative contributions consider mixtures of Gaussians [11, 13, 27] where, unfortunately, the increased flexibility comes at the cost of loosing interpretability and parsimony.

To balance these aspects in [2] we introduced a convex mixture regression model that interpolates two extremal densities $f_0(y)$ and $f_\infty(y)$ through a single monotone increasing function $\beta(x) \in [0, 1]$, which induces a flexible, yet interpretable and parsimonious, characterization of the conditional density of y given x via

$$f_x(y) = \{1 - \beta(x)\}f_0(y) + \beta(x)f_\infty(y), \quad x \geq 0. \quad (1)$$

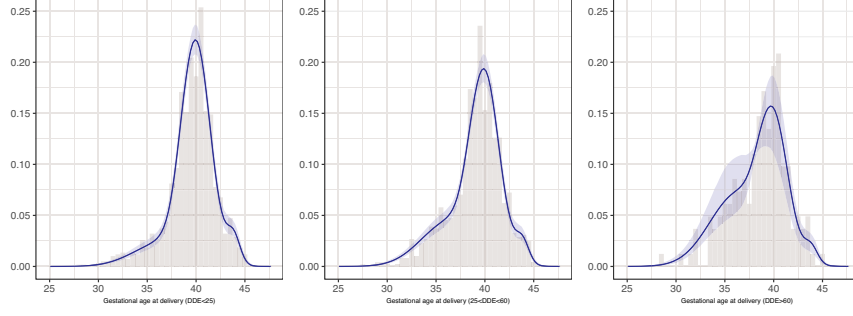


Fig. 1 Histograms of the gestational age at delivery for different (binned) values of DDE (gray) and posterior mean conditional densities under the convex mixture regression of [2].

The formulation in (1) has several appealing characteristics. First it induces a flexible model, generalizing the classical dose-response models for quantitative outcomes, with sufficient “structure” thus allowing to gain efficiency in estimation. This can be partially appreciated looking at the results reported in Figure 1 and in the detailed simulation study reported in [2]. Second, in contrast with the current trend of having black-box algorithms producing non interpretable results, our formulation improves interpretability as it can be shown that classical risk measures—such as the additional risk [14] for continuous outcomes—are proportional to the $\beta(x)$ function. Third, our Bayesian implementation naturally accounts for uncertainty in estimating benchmark dose BMD_q —defined as the dose corresponding to a 100q% risk of an adverse health event above that for unexposed individuals— and the more conservative benchmark doses lower bound (BMDL_q) without requiring asymptotic approximations. Fourth, it can be generalized to model health outcome y taking values on different sample spaces example—see for example [9].

2.2 Going beyond asymptotic extreme value theory: a Bayesian hierarchical extreme value model

In this section I deal with a classical topic in extreme value analysis [5], i.e. inference on the distribution of yearly maxima. Our specific motivating application is related to the estimation of yearly maxima of daily rainfall. Despite extreme value theory for this type of application is not novel, the data sets that are nowadays available drastically changed in the last decade due to the modern advances in remote sensing technologies. On one side, remote sensing increased the availability of spatial rainfall data, on the other side, these time series are naturally shorter as the technology is very recent. The result is that we have many spatially related and few temporally related observations. Our goal is to model short time series—*small*

data?—introducing external input on the analysis based on prior information for similar spatial locations by means of a Bayesian approach.

To set up the notation, consider to model the distribution of the maximum Y of an independent sequence of random variables X_1, \dots, X_n with common cumulative distribution function $F(\cdot; \theta)$ and $\theta \in \Theta$ and unknown parameter. Under these settings the distribution of Y is simply

$$\text{pr}_\theta(Y < y) = \text{pr}_\theta(X_1 < y, \dots, X_n < y) = F(y; \theta)^n. \quad (2)$$

A classical approach exploits the celebrated Fisher-Tippett Gnedenko theorem [8, 10] that states that for $n \rightarrow \infty$ the distribution in (2)—after a suitable normalization of Y —belongs to the family of the generalized extreme value (GEV) distribution [31], having cumulative distribution function equal to

$$F_{GEV}(y | \mu, \sigma, \xi) = \exp \left\{ - \left[1 + \frac{\xi}{\sigma} (y - \mu) \right]^{-1/\xi} \right\}.$$

Despite the elegant theoretical justification of this classic approach many limitations are actually present. First, the asymptotic argument. Trivially, in modelling year maxima of daily rainfall, n is at most 365 but also in general the rate of convergence to the theoretical extreme value distributions is not known and the number of events per block may be often not large enough for the asymptotic argument to hold [16]. Second, the assumption of no inter-block variability. In many applications, including our motivating examples we observe natural fluctuations in the intensity of the events. For these reasons and exploiting the natural hierarchical structure of the data generating process in the working paper by Zorzetto, Canale, and Marani [34] we develop a hierarchical Bayesian extreme value model that avoids the asymptotic argument, accounts for possible inter annual variability in the magnitude of the events, and, thanks to a careful prior elicitation, leads to more precise estimates than standard frequentist and Bayesian approaches. The ideas briefly summarized here are largely inspired by the seminal works [21, 33, 35].

Let n_j be the number of events over a given time period, or block j with $j = 1, \dots, J$ and x_{ij} the magnitude of the i -th event within the j -th time period with $i = 1, \dots, n_j$. The n_j events occurring within a block are assumed to be independent and identically distributed with common parametric cdf $F(\cdot; \theta_j)$ with θ_j a—possible multivariate—unknown parameter. Our proposal consists in eliciting the hierarchical model reported in Figure 2 where the n_j are realizations of random variables with common probability distribution $p(n; \lambda)$, and the latent θ_j are also realizations of a random variable with common probability density function $g(\cdot; \eta)$. Conditionally on n_j and θ_j , the single x_{ij} has pdf $f(x_{ij}; \theta_j)$. Prior distributions for λ and η complete the model's specification. The posterior distribution is approximated via Markov Chain Monte Carlo sampling.

Preliminary results show that the proposed approach leads to better estimates of the cdf of Y than classical or Bayesian implementations of the GEV approach and of the peak over threshold [6] approach. In addition the variability of such estimates—

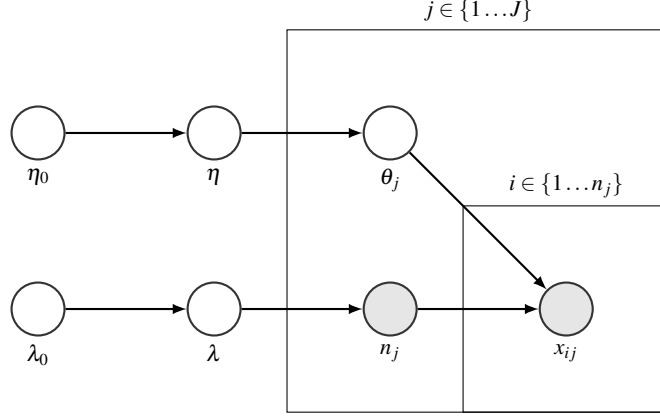


Fig. 2 Hierarchical structure of the model for the daily rainfall. Grey dots represent observed variables.

quantified in terms of Bayesian credible bands around the cdf—is smaller, leading to a more precise characterization of the probability of extreme events.

2.3 Sparsity for infinite dimensional data: a locally non-concurrent functional model for functional time series

In this section I briefly describe a model for a complex type of data that are nowadays routinely collected in many sectors. Specifically I am considering high-resolution data which can be modelled as smooth functions (e.g., curves or surfaces) that constitute the basis of functional data analysis, an emerging field of statistics where the single datum is not a number or vector but a function belonging to a suitable separable Hilbert space [12, 25, 26].

As motivating application, consider the problem of price prediction in modern energy markets and specifically its prediction by means of functional time series of demand and supply curves similarly to [3, 4, 15, 29]. In the latter contributions we studied different models for analyzing time series of functions, subject to equality and inequality constraints at the two edges of the domain, respectively, such as daily demand and offer curves.

In broader settings, consider a general functional regression situation where $y_i(s)$ is the functional response in L^2 for the i th observation and $x_i(s)$ is a functional regressor. The statistical literature proposes two different extreme solutions to deal with this situation. A first solution, called *concurrent functional model*, assumes that the relation between the functional response and the functional predictors is

$$y_i(s) = x_i(s)\phi(s) + e_i(s), \quad (3)$$

where $\phi(s)$ is a functional regression coefficient, and $e_i(s)$ are functional independent errors. The model is concurrent since the value of the functional response y_i at the domain point $s \in S$ depends only on the value of the functional regressors evaluated at the same $s \in S$. On the other extreme, there is the *non-concurrent functional model* that allows $y_i(s)$ to depend from the functional regressor entirely, and specifically

$$y_i(s) = \int x_i(t) \psi(t, s) dt + e_i(s), \quad (4)$$

where $\psi(t, s)$ is a kernel function which determines the impact of the function x_i evaluated at domain point $t \in S$ on $y_i(s)$.

Both specifications have pros and cons. The first is simple but there are cases in which it is not reasonable to assume that the value of $y_i(s)$ depends on the functional regressors in $s \in S$ only. The non-concurrent model, on the other side, offers great flexibility but this flexibility comes at the price of increasing complexity—both in terms of interpretation and computation.

In the working paper by Bernardi, Canale, and Stefanucci [1] we fill the gap between the concurrent and the non-concurrent models proposing a hybrid approach that matches the interpretability of the concurrent model with the flexibility of the non-concurrent model. We achieve this inducing a sparse and smooth block structure in the Hilbert-Schmidt coefficient ψ as sketched in what follows. Going back to our motivating application, assume that the response is y_t the value at day t (say today) of a demand curve while the functional covariate is y_{t-1} , its lag 1 observation (yesterday demand). Knowing that the impact of yesterday demand on today demand involves only some parts of their domain is certainly of dramatic interest to better understand the market dynamics.

To achieve this local sparsity structure, first represent both y_i and x_i via basis expansions, say $\{\phi_l(s), l = 1, 2, \dots, M\}$ and $\{\varphi_m(t), m = 1, 2, \dots, M\}$, to represent all the quantities involved in model (4), as

$$\begin{aligned} y_i(s) &= \sum_l \alpha_{il} \phi_l(s) = \Phi(s)^T \alpha_i \\ x_i(t) &= \sum_m \beta_{im} \varphi_m(t) = \boldsymbol{\varphi}(t)^T \beta_i. \end{aligned}$$

A similar representation can be provided for the kernel ψ , simply exploiting a tensor product expansion and defining

$$\psi(t, s) = \sum_l \sum_m \psi_{lm} \varphi_m(t) \phi_l(s), \quad (5)$$

or alternatively

$$\begin{aligned} \psi(t, s) &= (\varphi_1(t), \dots, \varphi_M(t)) \begin{pmatrix} \psi_{11} & \dots & \psi_{1M} \\ \vdots & \ddots & \vdots \\ \psi_{M1} & \dots & \psi_{MM} \end{pmatrix} \begin{pmatrix} \phi_1(s) \\ \vdots \\ \phi_M(s) \end{pmatrix} \\ &= \boldsymbol{\varphi}(t)^T \boldsymbol{\Psi} \boldsymbol{\Phi}(s). \end{aligned} \quad (6)$$

The desired local sparsity can be obtained by suitable defining the elements involved in equation (6). For example, if we opt for B-splines [7] of order d a sufficient condition to get $\psi(t, s) = 0$ is to have a $d \times d$ block of the matrix Ψ of zeroes. This can be obtained borrowing ideas from the group lasso regression [32].

Acknowledgements

All the materials presented in this short paper comes from collaborative projects and I want to thank all my coauthors: Daniele Durante and David Dunson for Section 2.1, Enrico Zorzetto and Marco Marani for Section 2.2, and Marco Stefanucci and Mauro Bernardi for Section 2.3. Comments from Saverio Ranciati and Bernardo Nipoti are also gratefully acknowledged. The projects discussed in Sections 2.1–2.2 are supported by the University of Padova under the STARS Grants programme BNP-CD.

References

- [1] M. Bernardi, A. Canale, and M. Stefanucci. Locally non-concurrent functional regression models. In preparation, 2019.
- [2] A. Canale, D. Durante, and D. B. Dunson. Convex mixture regression for quantitative risk assessment. *Biometrics*, 74:1331–1340, 2018.
- [3] A. Canale, M. Ruggiero, et al. Bayesian nonparametric forecasting of monotonic functional time series. *Electronic Journal of Statistics*, 10(2):3265–3286, 2016.
- [4] A. Canale and S. Vantini. Constrained functional time series: Applications to the Italian gas market. *International Journal of Forecasting*, 32(4):1340–1351, 2016.
- [5] S. Coles. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.
- [6] A. C. Davison and R. L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3):393–425, 1990.
- [7] C. De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.
- [8] R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge University Press, 1928.
- [9] L. Galtarossa and A. Canale. A convex mixture model for binomial regression. In *Book of Short Papers SIS 2019*, 2019.

- [10] B. Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, pages 423–453, 1943.
- [11] B. He, M. Chen, L. Song, and D. Wang. Mixture normal models in which the proportions of susceptibility are related to dose levels. *Acta Mathematicae Applicatae Sinica, English Series*, 26(3):463–472, 2010.
- [12] L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer New York, 2012.
- [13] B. S. Hwang and M. L. Pennell. Semiparametric Bayesian joint modeling of a binary and continuous outcome with applications in toxicological risk assessment. *Statistics in Medicine*, 33(7):1162–1175, 2014.
- [14] R. L. Kodell and R. W. West. Upper confidence limits on excess risk for quantitative responses. *Risk Analysis*, 13(2):177–182, 1993.
- [15] G. Kon Kam King, A. Canale, and M. Ruggiero. Bayesian functional forecasting with locally-autoregressive dependent processes. *Bayesian Analysis*, 2018.
- [16] D. Koutsoyiannis. Statistics of extremes and estimation of extreme rainfall: I. theoretical investigation. *Hydrological sciences journal*, 49(4), 2004.
- [17] B. Landman and M. J. Cardoso. First international workshop on deep learning fails. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 43–94. Springer, 2018.
- [18] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436 EP –, 05 2015.
- [19] J. Lehman, J. Clune, D. Misevic, C. Adami, J. Beaulieu, P. J. Bentley, S. Bernard, G. Belson, D. M. Bryson, N. Cheney, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *arXiv preprint arXiv:1803.03453*, 2018.
- [20] M. P. Longnecker, M. A. Klebanoff, H. Zhou, and J. W. Brock. Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. *The Lancet*, 358(9276):110–114, 2001.
- [21] M. Marani and M. Ignaccolo. A metastatistical approach to rainfall extremes. *Advances in Water Resources*, 79:121–126, 2015.
- [22] G. Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.
- [23] J. Markoff. Scientists see promise in deep-learning programs. *The New York Times*, 2012.
- [24] W. W. Piegorsch and A. J. Bailer. *Analyzing Environmental Data*. John Wiley & Sons, 2005.
- [25] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2002. Methods and case studies.
- [26] J. O. Ramsay and B. W. Silverman. *Functional data analysis*, volume 2nd Ed. Springer, New York, 2005.

- [27] M. Razzaghi and R. L. Kodell. Risk assessment for quantitative responses using a mixture model. *Biometrics*, 56(2):519–527, 2000.
- [28] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [29] J. Rossini and A. Canale. Quantifying prediction uncertainty for functional-and-scalar to functional autoregressive models under shape constraints. *Journal of Multivariate Analysis*, 170:221–231, 2019.
- [30] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.
- [31] R. Von Mises. La distribution de la plus grande de n valeurs. *Rev. math. Union interbalcanique*, 1(1), 1936.
- [32] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.
- [33] E. Zorzetto, G. Botter, and M. Marani. On the emergence of rainfall extremes from ordinary events. *Geophysical Research Letters*, 43(15):8076–8082, 2016.
- [34] E. Zorzetto, A. Canale, and M. Marani. A hierarchical Bayesian model for extreme value analysis. In preparation, 2019.
- [35] E. Zorzetto and M. Marani. Downscaling of rainfall extremes from satellite observations. *Water Resources Research*, 2018.

Smart Statistics: concept, technology and service

David John Hand, Maurizio Vichi

Abstract Human Computer-Mediated-Communication (CMC) and the Internet of Things, (IoT) a non-human CMC, have changed the styles of communication producing a great variety and volume of data at high velocity. In CMC human communication occurs through two or more “smart devices” having specific formats, such as instant messaging, email, chat rooms, online forums, social network services and text messaging. Communication, with CMC, happens in real time (synchronous), or at differed time (asynchronous), with parties not communicating at the same time. The by-product of CMC is the “*datafication*” of different aspects of the life of citizens. Social media platforms tend to broadcast continuously and show communication among people on all possible topics and phenomena relating to their lives. Society collectively accumulates data on massive amounts of its behaviours. On the other hand, IoT produces communication between things and a datafication of actions and services for which they are used. In general, we can say that CMC and IoT produce organic data, i.e., rough data not ready to be statistically analysed. Information/knowledge are different from organic data. To extract something useful from the data, sophisticated statistical methodologies, and algorithms, and other tools are required. Moreover, the velocity of the flow of incoming data often means that their analysis needs to be done in real time. The integration of the data with the statistical methodologies necessary to transform these data into trusted information/knowledge produces what can be called “smart statistics”. This is a multidimensional concept connected with three relevant topics: new styles of communication and datafication in the Society; the change of Statistical Sciences due to internet and the new technologies; the use of “smart systems” as services for the Society. This paper looks at the potential implications of smart statistics. We wish to define the term of Smart Statistics by briefly describing the topics that characterise this multidimensional concept.

Key words: Smart Statistics, Computer-Mediated-Communication, Internet of Things

¹

David John Hand, Imperial College, London;

Maurizio Vichi, Sapienza University, Rome; maurizio.vichi@uniroma1.it:

Tavola rotonda “Smart ageing: lunga vita attiva, salute e nuove tecnologie”

ORGANIZZATORI: A. Rosina, G. Rivellini, M.P. Vitale

MODERATORE: Francesco Cancellato (Linkiesta).

RELATORI: Stefano Campostrini (Università Ca Foscari), Maria Grazia Carrozza (Scuola Superiore Sant'Anna), Annibale Delia (Comune di Milano), Alessandra De Rose (Sapienza Università di Roma e Laboratorio LoLA - Longevity and Ageing)

Aumento della longevità e innovazione tecnologica sono due grandi trasformazioni in atto destinate a cambiare profondamente il nostro modo di vivere, stare in relazione, produrre e promuovere benessere. Non si tratta solo di quantità in aumento (di anni di vita e di popolazione in età anziana) ma di qualità da aggiungere a fasi della vita in grande mutamento. La quantità va anzi considerata una opportunità per investire sulla qualità. La silver economy destinata ad avere un impatto crescente sull'economia, con crescente ruolo delle nuove tecnologie ad ogni livello. La crescita della popolazione anziana e le politiche che favoriscono un invecchiamento attivo tendono a favorire la domanda di mantenimento in buona salute, ma più in generale di buona qualità della vita, con possibilità di continuare a vivere in autonomia nella propria abitazione, senza sentirsi isolati e potendo contare su adeguati servizi di assistenza e su solide reti di aiuto sociale. Una risposta a tale domanda arriva anche dalla sperimentazione e dagli investimenti in ricerca e sviluppo. Un esempio quello della progettazione di abitazioni intelligenti (smart home), caratterizzate da un ambiente monitorato e sicuro, grazie alla tecnologia digitale, che si adatta alle esigenze di chi la vive e ne facilita le attività. L'espansione di prodotti e servizi in questa direzione vede come protagonista l'innovazione prodotta dalle nuove generazioni. Inoltre l'evoluzione della smart home ha ricadute potenziali su tutte le abitazioni, generando quindi benefici estendibili a tutti i cittadini. Lo stesso vale per i trasporti e vari altri aspetti della vita nelle presenti e future smart cities. Una lunga vita attiva di qualità si costruisce a partire dalle età più giovani, attraverso formazione continua, sviluppo di life skills, competenze digitali, smart working, sistemi esperti in grado di accompagnare i passaggi nelle varie fasi della vita e di conciliare varie dimensioni di vita. Quanto stiamo davvero preparando le nuove generazioni ad una lunga vita attiva di successo? Quali rischi e opportunità le nuove tecnologie possono offrire?



Section 2. Invited Papers

Demography in the Digital Era: New Data Sources for Population Research

Demografia nell'era digitale: nuovi fonti di dati per gli studi di popolazione

Diego Alburez-Gutierrez, Samin Aref, Sofia Gil-Clavel, André Grow, Daniela V. Negraia, Emilio Zagheni

Abstract The spread of digital technologies and the increased access to the internet has contributed to the production and accumulation of unprecedented quantities of data about human behavior. Demographers, who have a long-standing interest in issues related to data and data quality, are in an ideal position to make sense of this new information. This paper discusses three ways in which the Data Revolution has created novel sources of data for demographic research. It discusses the unique technical and ethical challenges posed by these data sources and the opportunities they provide for understanding historical and contemporary demographic dynamics around the world.

Abstract *La diffusione di tecnologie digitali e la crescita nell'accesso ad internet hanno contribuito ad una produzione ed accumulo senza precedenti di dati sul comportamento umano. I demografi, che hanno avuto un interesse di lunga data su questioni relative a dati e qualità dei dati, sono in una posizione ideale per estrarre significato da queste nuove informazioni. Questo articolo discute tre modi in cui la 'Data Revolution' ha creato nuove fonti di dati per la ricerca demografica. L'articolo discute le sfide tecniche ed etiche create da questi dati e le opportunità che offrono per comprendere le dinamiche demografiche storiche e contemporanee.*

Key words: demography, digital data, social media, data revolution

Diego Alburez-Gutierrez, Max Planck Institute for Demographic Research (MPIDR), e-mail: alburezgutierrez@demogr.mpg.de
Samin Aref, MPIDR, e-mail: aref@demogr.mpg.de
Sofia Gil-Clavel, MPIDR, e-mail: gil@demogr.mpg.de
André Grow, MPIDR, e-mail: grow@demogr.mpg.de
Daniela V. Negraia, MPIDR, e-mail: negraia@demogr.mpg.de
Emilio Zagheni, MPIDR, e-mail: zagheni@demogr.mpg.de

1 The Data Revolution: A New Data Paradigm in Demography?

Demography, the systematic study of population dynamics and the causes and consequence of compositional changes in populations, has always been a data-driven discipline. Administrators have used censuses to count (and tax) populations since ancient times. In modern societies, an interest in data characterized the development of the discipline of demography. For example, John Graunt identified London’s 16th century ‘Bills of Mortality’ as a potential source of data for demographic analysis, ultimately resulting in the creation of life tables. We argue that demography is at the gates of a new data paradigm defined by the increased availability of population data produced or made available by digital technologies and the internet [6, 8]. The shift is part of the Data Revolution, the process through which the transition from analogue to digital electronic technologies has resulted in the accumulation of vast amounts of individual-level data (see Figure 1 for an illustration). The spread of the internet, the World Wide Web, and the Internet of Things, have accelerated this process, producing unprecedented data on society and human behavior [29].

This paper presents three innovative sources of data that have been made possible by the Data Revolution and explores their potential for conducting groundbreaking demographic research. First, digitization has helped improve access to existing data, such as censuses and population registers [33, 20], and bibliometric databases [12, 27]. Similarly, the advent of online peer-to-peer collaboration has created new resources, such as massive online genealogical databases [21], that can be used for studying intergenerational demographic processes. Second, demographers can now analyze digital traces left by internet users in platforms like Twitter [35] and Facebook (FB) [15] to study population dynamics. Finally, the Data Revolution has created new opportunities for collecting primary data using devices connected to the internet. Examples discussed in this paper include online surveys [7, 1], apps for registering time-use data [28], and internet advertising platforms [37, 9].

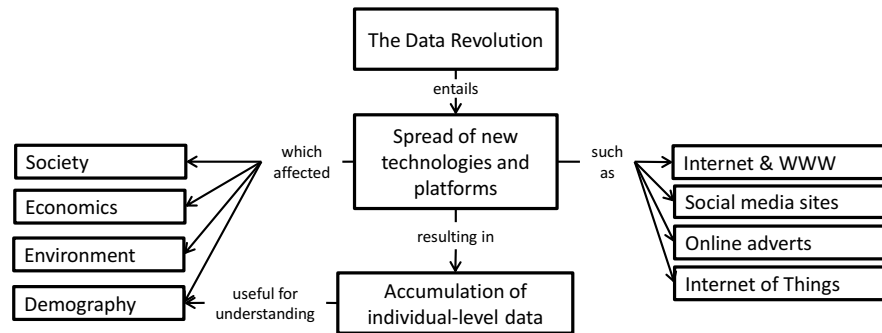


Fig. 1 The Data Revolution and new sources of data for demographic analysis.

1.1 Digitized and Crowd-sourced Data

Demographers and statistical agencies were quick to recognize the importance of digitizing paper-based demographic data. The digitization of censuses and population registers was pioneered by the Integrated Public Use Microdata Series (IPUMS), which now hosts the world's largest collection of demographic microdata.¹ In time, digitization enabled the creation of crucial data repositories for demographic research (e.g. the Human Mortality and the Human Fertility Databases² or digital national population registers). Nordic registers, for example, have been used to study intergenerational processes in fertility [22], health [5], mortality [4], and migration [33]. Most of the existing research focuses on Europe, but researchers increasingly acknowledge the potential of other population registers for conducting demographic research (e.g. East Asia [13] or North America [18]).

Bibliometric databases, such as Scopus [2], Web of Science [27], and Dimensions [32], are other examples of digitized sources with potential for demographic research. These databases contain data on millions of scientific publications produced each year, including author affiliation and addresses. Affiliation data can be used for analyzing scientific collaboration and mobility of researchers across countries [24, 3, 12]. Yet, using these data sources for migration research has limitations which require a careful interpretation of the results [2, 27]. Changes on author affiliation, for example, are not a perfect proxy for mobility since conducting and publishing research can be a lengthy process. Migration of researchers is likely to be underestimated because some movements are not represented in publications indexed in bibliometric databases. This calls for future research integrating bibliometric data with complementary data sources to resolve some of the methodological issues. Despite these limitations, bibliometric data sources offer substantial benefits [12, 27] compared to traditional data sources like surveys. These resources make research on migration of research-active scientists more cross-disciplinary, scalable, longitudinal, contemporary, and comprehensive.

Demographic data can also be crowd-sourced. Platforms like Geni.com and WikiTree have allowed thousands of amateur genealogists to collaborate in building large-scale online genealogical databases such as the *Familinx* database, which includes 86 million individual records from around the globe, with data that go back as far as the 17th century [21]. This particular database was scraped from Geni.com, a collaborative social network that allows users to find and verify family relations. Online genealogies are a promising resource because they cover long historical periods and are not restricted by national boundaries - on the downside, they are not representative samples and underrepresent Low- and Middle-Income countries (LMIC). Despite their potential, sound demographic research using these data is still missing, including methodologies for addressing systematic biases and generalizing the findings to larger populations [16].

¹ www.ipums.org; www.international.ipums.org, accessed 28.02.2019.

² www.mortality.org; www.humanfertility.org, accessed 28.02.2019.

1.2 Digital Traces from Social Media

About half of the world's population are active internet users and many use social media platforms like FB and Twitter.³ Demographic information on the users of these platforms can be used to perform demographic research in a timely manner. Social media data can also be used to study populations that would otherwise remain entirely out of reach [26]. Researchers can access FB and Twitter data using the platforms' Application Programming Interfaces (API), some of which have been designed for advertising purposes. The FB Marketing API gives access to aggregated population data (e.g., the number of FB users by sex and age in a given country who share certain interests), but not individual-level user data. Unfortunately, FB does not provide much detail about how these aggregate figures are estimated (e.g., how users are classified according to their interests, behavior, and demographic characteristics). Twitter, in addition to a Marketing API for aggregate-level summary information, allows researchers to query individual-level data from 'public tweets' (i.e. tweets not protected by the user).⁴ Still, researchers can access information that users have agreed to share, including text and images from tweets, user names, and tweet locations. Having access to individual-level Twitter posts gives researchers the freedom to design and test different models and algorithms using primary data.

Previous studies have collected data using APIs to study contemporary social and demographic processes. FB data have been used to study access to digital technologies [15, 17], immigrant cultural assimilation [14], and to estimate migrant stocks [37]. Twitter data have been used to study migration flows [35], and monitor population health [10] and natural disasters [19]. The use of the FB and Twitter data has clear advantages, but also important drawbacks. A notable limitation is that social media data are generally not representative of the entire population. Recent studies have attempted to overcome this limitation by combining social media data, statistical models, and representative surveys [36, 37]. Another limitation is the lack of individual-level demographic data for Twitter users. Studies have addressed this limitation by using pattern recognition techniques to infer the demographic characteristics of users [34]. Nevertheless, there are clear benefits in using this new source of data. For instance, demographers and sociologists have been able to reach and study new populations, while statisticians and computer scientists have had the opportunity to test new models and algorithms. These examples show how the internet has created research opportunities that were unimaginable when social networking platforms were initially conceived, over 20 years ago.

³ www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx, accessed 28.02.2019.

⁴ help.twitter.com/en/safety-and-security/public-and-protected-tweets, accessed 28.02.2019.

1.3 New opportunities for collecting primary data

The Data Revolution has also created new opportunities for collecting primary data via the internet. Several studies have recruited participants for online surveys using social networking sites (e.g., FB and Twitter) and online labor markets (e.g., Amazon Mechanical Turk and Craigslist) [7]. Such platforms tend to have wide reach and often allow the targeting of individuals based on specific demographic characteristics, interests, and behaviors. This makes them attractive for both drawing convenience samples and recruiting members of hard-to-reach populations, usually at a lower cost than would be possible with traditional probability samples [1]. Of the existing platforms, FB has been the most popular, arguably because it (currently) has the largest number of users and the widest international coverage, and because it provides detailed information about user characteristics that can be used for targeting participants [7, 9]. In the existing studies using FB, recruitment usually takes place via ads that can be shown to users at various places of a webpage. Such ads consist of one or more pictures accompanied by a short study description and a link to an external site hosting the survey (see [26] for an example).

In addition to new platforms for survey research, internet-enabled devices (e.g., mobile phones and activity trackers) can revolutionize current research practice. One example comes from the area of time-use research. The ways in which people use their time (i.e., the quantity and quality of time; whether they spend it alone or interacting with other people or with machines) has implications for their health and wellbeing. Information about people's time use and wellbeing has mainly been collected using (1) recall or (2) real-time techniques. In recall techniques, interviews are typically conducted over the phone or via paper-and-pencil self-administered diaries, where respondents report back about what they did during that day or the previous day, the duration of each activity, where they were, and who they interacted with. This method affords coverage and detail of what participants did during the previous day and the sequence in which activities took place, but it is not very precise in estimating the hour and the minute in which an activity took place [25]. Furthermore, such recall diaries can be lengthy and burdensome for the respondent, which is why most national representative recall-diary surveys are cross-sectional and only cover one or two diary-days for each respondent. 'Real-time' techniques that rely on applications or instant messages received on personal mobile phone have the potential to capture what people are doing 'right now', and are likely to provide a more precise picture of the exact activity [11, 28]. Such techniques also have the advantage of being faster and less fatiguing for the respondent, allowing longitudinal or repeated measures. Additionally, data collection via cellphone applications could provide much needed insight into what people from various socioeconomic contexts and political regimes do during day-to-day life and how those activities are then linked to various measures of wellbeing. Nevertheless, assessing time-use and subjective wellbeing in real time has its own reliability and validity challenges, stemming particularly from the fact that asking respondents to evaluate their current behavior and/or emotional wellbeing 'right now', may change the very behavior and/or emotions we are trying to measure [23, 31].

2 What's next for Demography?

This paper highlighted new opportunities for demographic research created by the Data Revolution. The review of new data sources, however, is not exhaustive and researchers will continue to find new ways of making sense of our social world with the help of the internet and electronic devices. This concluding section considers the unique technical and ethical challenges of digital data and discusses how addressing them can contribute to the advancement of the demographic discipline.

Demographers using digital data face particular issues related to access, representativity, and ethics. Researchers often 'depend on the kindness of strangers' for accessing data since internet companies, unlike governments, are not obliged to share data from their platforms. This creates uncertainty as the conditions of access may change in the future. There are important attempts to address this issue. The Opal Project, for example, has proposed protocols for private companies to willingly share anonymized data on a regular basis to inform public policy and academic research.⁵ Furthermore, digital sources are rarely representative of larger populations in the way that randomized surveys are (even if, as this paper has shown, digital technology can enhance the collection of primary survey data). Coverage can also be an issue, as access to the internet is more restricted in LMIC. Nevertheless, digital trace data can be used to show some of these global inequalities in access to digital technologies [15, 17].⁶ The issue has motivated research on generalizing from non-representative samples to larger populations [36, 37]. This is a promising area of methodological development with wide applications, especially as survey response rates continue to decline around the world. The availability of online data has also led researchers to think long and hard about data security, privacy and informed consent in the digital era [30]. Ethical considerations must be a primary concern when designing demographic studies using digital or internet data. Social scientists need to adhere to ethical and transparent research practices, particularly as the privacy of users is constantly threatened in the online world [38].

Finally, it is important to note that while innovative sources of data provide exciting opportunities for new research, they are unlikely to make 'traditional' demographic sources obsolete in the near future (e.g., surveys, censuses). Rather, the Data Revolution has the potential to complement and augment these existing data sources. Traditional population data, for example, are crucial for identifying systematic bias in online sources and calibrating estimates made from these data [37]. Social media data can be used to estimate important demographic measures in contexts where traditional survey data are not available. The Data Revolution has already changed the way we do demography, as evidenced by the digitization of historical censuses and populations registers, and the creation of large-scale and open-access repositories of demographic data. The pace of this changes is likely to increase in the future as more researchers engage in ground-breaking research using digital data sources.

⁵ www.opalproject.org/, accessed 28.02.2019.

⁶ This work has resulted in efforts to 'nowcast' the digital gender gap in internet and mobile access using real-time big data: <https://www.digitalgendergaps.org/>, accessed 28.02.2019.

References

1. Antoun, C., Zhang, C., Conrad, F.G., Schober, M.F.: Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field Methods* **28**(3), 231–246 (2016)
2. Appelt, S., van Beuzekom, B., Galindo-Rueda, F., de Pinho, R.: Which factors influence the international mobility of research scientists? In: A. Geuna (ed.) *Global Mobility of Research Scientists*, pp. 177–213. Academic Press, San Diego (2015)
3. Aref, S., Friggens, D., Hendy, S.: Analysing scientific collaborations of New Zealand institutions using Scopus bibliometric data. In: *Proceedings of the Australasian Computer Science Week Multiconference*, p. 49. ACM (2018)
4. Baranowska-Rataj, A., Barclay, K., Kolk, M.: The effect of number of siblings on adult mortality: Evidence from Swedish registers for cohorts born between 1938 and 1972. *Population Studies* **71**(1), 43–63 (2017)
5. Barclay, K.J., Kolk, M.: The long-term cognitive and socioeconomic consequences of birth intervals: a within-family sibling comparison using Swedish register data. *Demography* **54**(2), 459–484 (2017)
6. Billari, F.C., Zagheni, E.: Big data and population processes: A revolution? In: A. Petrucci, R. Verde (eds.) *Proceedings of the Conference of the Italian Statistical Society*, pp. 167–178. Firenze University Press (2017)
7. Boas, T.C., Christenson, D.P., Glick, D.M.: Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. *Political Science Research and Methods* pp. 1–19 (2018)
8. Cesare, N., Lee, H., McCormick, T., Spiro, E., Zagheni, E.: Promises and pitfalls of using digital traces for demographic research. *Demography* **55**(5), 1979–1999 (2018)
9. Chu, J.L., Snider, C.E.: Use of a social networking web site for recruiting Canadian youth for medical research. *Journal of Adolescent Health* **52**(6), 792–794 (2013)
10. Cocos, A., Fiks, A.G., Masino, A.J.: Deep learning for pharmacovigilance: Recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *Journal of the American Medical Informatics Association* **24**(4), 813–821 (2017)
11. Csikszentmihalyi, M., Larson, R.: Validity and reliability of the experience-sampling method. In: *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*, pp. 35–54. Springer, Dordrecht (2014)
12. Czaika, M., Orazbayev, S.: The globalisation of scientific mobility, 1970–2014. *Applied Geography* **96**, 1–10 (2018)
13. Dong, H., Campbell, C., Kurosu, S., Yang, W., Lee, J.Z.: New sources for comparative social science: Historical population panel data from East Asia. *Demography* **52**(3), 1061–1088 (2015)
14. Dubois, A., Zagheni, E., Garimella, K., Weber, I.: Studying migrant assimilation through Facebook interests. In: S. Staab, O. Koltsova, D.I. Ignatov (eds.) *Social Informatics, Lecture Notes in Computer Science*, pp. 51–60. Springer International Publishing (2018)
15. Fatehkia, M., Kashyap, R., Weber, I.: Using Facebook ad data to track the global digital gender gap. *World Development* **107**, 189–209 (2018)
16. Fire, M., Elovici, Y.: Data mining of online genealogy datasets for revealing lifespan patterns in human population. *ACM Trans. Intell. Syst. Technol.* **6**(2), 28:1–28:22 (2015)
17. Garcia, D., Kassa, Y.M., Cuevas, A., Cebrian, M., Moro, E., Rahwan, I., Cuevas, R.: Analyzing gender inequality through large-scale Facebook advertising data. *Proceedings of the National Academy of Sciences* **115**(27), 6958–6963 (2018)
18. Gauvin, H., Moreau, C., Lefebvre, J.F., Laprise, C., Vézina, H., Labuda, D., Roy-Gagnon, M.H.: Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. *European Journal of Human Genetics* **22**(6), 814–821 (2014)
19. Ghahremanlou, L., Sherchan, W., Thom, J.A.: Geotagging Twitter messages in crisis management. *The Computer Journal* **58**(9), 1937–1954 (2015)

20. Hall, P.K., McCaa, R., Thorvaldsen, G., Group, I.M.A. (eds.): Handbook of international historical microdata for population research: A project of IMAG, The International Microdata Access Group. Minnesota Population Center, Minneapolis, Minn (2000)
21. Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., Bhatia, G., MacArthur, D.G., Price, A.L., Erlich, Y.: Quantitative analysis of population-scale family trees with millions of relatives. *Science* **360**(6385), 171–175 (2018)
22. Kolk, M.: Multigenerational transmission of family size in contemporary Sweden. *Population Studies* **68**(1), 111–129 (2014)
23. Ludwigs, K., Lucas, R., Burger, M., Veenhoven, R., Arends, L.: How does more attention to subjective well-being affect subjective well-being? *Applied Research in Quality of Life* **13**(4), 1055–1080 (2018)
24. Moed, H.F., Halevi, G.: A bibliometric approach to tracking international scientific migration. *Scientometrics* **101**(3), 1987–2001 (2014)
25. Phipps, P.A., Vernon, M.K.: Twenty-four hours: an overview of the recall diary method and data quality in the American time use survey. In: R.F. Belli, F.P. Stafford, D.F. Alwin (eds.) *Calendar and Time Diary: Methods in Life Course Research*, pp. 109–128. Sage Publications, Thousand Oaks (2009)
26. Pötzschke, S., Braun, M.: Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries. *Social Science Computer Review* **35**(5), 633–653 (2017)
27. Robinson-Garcia, N., Sugimoto, C.R., Murray, D., Yegros-Yegros, A., Larivière, V., Costas, R.: The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics* **13**(1), 50–63 (2019)
28. Roessger, K.M., Greenleaf, A., Hoggan, C.: Using data collection apps and single-case designs to research transformative learning in adults. *Journal of Adult and Continuing Education* **23**(2), 206–225 (2017)
29. Salathé, M., Bengtsson, L., Bodnar, T.J., Brewer, D.D., Brownstein, J.S., Buckee, C., Campbell, E.M., Cattuto, C., Khandelwal, S., Mabry, P.L., Vespignani, A.: Digital Epidemiology. *PLoS Computational Biology* **8**(7), e1002616 (2012)
30. Salganik, M.J.: Bit by bit: Social research in the digital age. Princeton University Press, Princeton (2018)
31. Sasaki, W., Nakazawa, J., Okoshi, T.: Comparing ESM timings for emotional estimation model with fine temporal granularity. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, pp. 722–725. ACM (2018)
32. Thelwall, M.: Dimensions: A competitor to Scopus and the Web of Science? *Journal of Informetrics* **12**(2), 430–435 (2018)
33. Thorvaldsen, G., Østrem, N.O.: Migration and the historical population register of Norway. *Journal of Migration History* **4**(2), 237–248 (2018)
34. Yildiz, D., Munson, J., Vitali, A., Tinati, R., Holland, J.A.: Using Twitter data for demographic research. *Demographic Research* **37**(46), 1477–1514 (2017)
35. Zagheni, E., Garimella, V.R.K., Weber, I., State, B.: Inferring international and internal migration patterns from Twitter data. In: Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion, pp. 439–444. ACM Press, Seoul, Korea (2014)
36. Zagheni, E., Weber, I.: Demographic research with non-representative internet data. *International Journal of Manpower* **36**(1), 13–25 (2015)
37. Zagheni, E., Weber, I., Gummadi, K.: Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review* **43**(4), 721–734 (2017)
38. Zuboff, S.: Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology* **30**(1), 75–89 (2015)

Stationarity of a general class of observation driven models for discrete valued processes

Stazionarietà di una classe generale di modelli observation-driven per processi a valori discreti

Mirko Armillotta, Alessandra Luati and Monia Lupparelli

Abstract A large variety of time series observation-driven models for binary and count data are currently used in different contexts. Despite the importance of stationarity and ergodicity to ensure suitable results, for many of these models stationarity is not yet proved. We specify a general class of observation-driven models for discrete valued processes, which encompasses the most frequently used models. Then, we show strict stationarity by means of Feller properties and establish easy-to-check stationarity conditions.

Abstract *Modelli observation-driven per serie storiche di dati binari e di conteggio sono correntemente utilizzati in diversi contesti. In alcuni casi, tuttavia, le proprietà di stazionarietà ed ergodicità non sono state dimostrate. In questo paper, viene specificata una classe generale di modelli observation driven per dati discreti, che comprende i modelli maggiormente utilizzati in letteratura. Tramite le proprietà di Feller, si derivano condizioni di stazionarietà semplici da verificare.*

Key words: Generalized linear ARMA, Time series of counts, Binary variables, Drift conditions

1 Introduction

Observation-driven models were originally introduced by Cox [2] and they have nowadays received new interest. There is an heterogeneous literature about such models for binary data [12, 17, 11] and for count data [3, 7, 9]; other general models

Mirko Armillotta, Alessandra Luati

Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, 40126, Bologna, e-mail: mirko.armillotta2@unibo.it, e-mail: alessandra.luati@unibo.it

Monia Lupparelli

Department of Statistics, Computer Science, Applications, University of Florence, Viale Morgagni 59, 50134, Florence, e-mail: monia.lupparelli@unifi.it

were introduced in [1] and [19]. More recently, various attempts have been done to study the probabilistic properties of these models. Stationarity and ergodicity were proved for a very general model in [5] and in [4], but these results do not directly apply to the models mentioned above; the results of [5] and [4] provide a basis for proofs which one needs to develop from time to time, depending on different models and specific distributions. In practical applications, directly applicable stationarity condition are needed to guarantee the reliability and validity of the results obtained. Strict stationarity results have been directly derived by [13] solely for the Generalized Autoregressive Moving Average (GARMA) model of [1].

Our contribution extends the argument of [13] and provides stationarity and ergodicity conditions directly verifiable and applicable for a class of observation-driven models that encompasses the models mentioned so far and for data coming from a large family of distributions.

In Section 2 we formulate the general framework, with some examples. In Section 3 we establish stationarity and ergodicity for the model. In Section 4 we apply the results to some specific models. In Section 5, concluding remarks and future developments are highlighted.

2 The framework

Let us consider the stochastic process $\{Y_n\}_{n \in \mathbb{N}}$ and the filtration $\mathcal{F}_{n-1} = \sigma(Y_s, s \leq n-1, X_0 = x)$, the information set up to time $n-1$ and the starting value for X_n . An observation-driven model for Y_n has the form

$$Y_n | Y_{0:n-1} \sim f(\cdot; \mu_n) \quad (1)$$

$$\mu_n = q_{\theta,n}(Y_{0:n-1}) \quad (2)$$

where $q_{\theta,n}$ is some function parametrized by θ and $f(\cdot; \mu_n)$ is a density (or mass) function whose dynamic is captured by μ_n ; usually, but not necessarily, this distribution is assumed to belong to the exponential family with μ_n as conditional expectation. We focus on models where the observation process $\{Y_n\}_{n \in \mathbb{N}}$ is integer-valued and find conditions under which there exists a stationary and ergodic version of it via Markov chain theory. However, since $\{Y_n\}_{n \in \mathbb{N}}$ is not itself a Markov chain, a classical approach is to prove the existence of a stationary ergodic process $\{Y_n\}_{n \in \mathbb{N}}$ as a function of an ergodic Markov chain $X = \{X_n\}_{n \in \mathbb{N}}$, on a state space S with σ -algebra \mathcal{F} and n -step transition kernel $P(x, A) = \mathbb{P}^n(X_n \in A | X_0 = x)$ for $A \in \mathcal{F}$ and starting from $X_0 = x$.

In the present case, the chain is specified as $X_n = g(\mu_n)$ where g is a bijective increasing function, the *link function*; an explicit formulation for (2) is defined as follows

$$g(\mu_n) = \alpha + \sum_{j=1}^k \gamma_j g(\mu_{n-j}) + \sum_{j=1}^p \phi_j h(Y_{n-j}) + \sum_{j=1}^q \theta_j \left[\frac{h(Y_{n-j}) - \bar{g}(\mu_{n-j})}{v_{n-j}} \right] \quad (3)$$

Stationarity of general class of observation driven models for discrete valued processes

where v_n is some scaling sequence. The function $h(Y_n)$ is called *y-link function* because it is applied only to the observations Y_n whereas $\bar{g}(\mu_n)$ is said *mean-link function* because it is applied only to the expected value μ_n . Both link functions are monotone and could be different from the link $g(\cdot)$. In general, it is useful to choose the mean-link function as follows:

$$\bar{g}(\mu_n) = E[h(Y_n) | \mathcal{F}_{n-1}], \quad (4)$$

so that $\varepsilon_n = h(Y_n) - \bar{g}(\mu_n)$ is a martingale difference sequence (MDS) and can be interpreted as a *prediction error*.

For sake of clarity we focus the attention on the first order model

$$g(\mu_n) = \alpha + \gamma g(\mu_{n-1}) + \phi h(Y_{n-1}^*) + \theta \left[\frac{h(Y_{n-1}^*) - \bar{g}(\mu_{n-1})}{v_{n-1}} \right], \quad (5)$$

where Y_n^* is some mapping of Y_n to the domain of $h(\cdot)$.

The general class of models (5) has a large flexibility in that it encompasses many time series models of interest. The GARMA model [1, 13] is easily obtained when $\gamma = 0$, by setting $g \equiv \bar{g} \equiv h$ and $v_n = 1$, such as, by equivalence of the three link functions and no scaling applied, one has

$$g(\mu_n) = \alpha + \phi g(Y_{n-1}^*) + \theta [g(Y_{n-1}^*) - g(\mu_{n-1})]. \quad (6)$$

Note that, in this case, $\varepsilon_n = g(Y_{n-1}^*) - g(\mu_{n-1})$ is a MDS only in the special case in which $g \equiv h$, the identity function.

The Binomial ARMA (BARMA) model, developed in [12, 17] is obtained by (5) when $\gamma = 0$, h is the identity ($\bar{g}(\mu_t)$ reduces to μ_t) and $Y_n^* = Y_n$. Then,

$$g(\mu_n) = \alpha + \phi Y_{n-1} + \theta [Y_{n-1} - \mu_{n-1}]. \quad (7)$$

Another promising branch of the literature has been developed by [16] and [3], under the name of Generalized Linear ARMA (GLARMA) models. This class is recovered here by setting $\phi = 0$ and h the identity,

$$g(\mu_n) = \alpha + \gamma g(\mu_{n-1}) + \theta \left[\frac{Y_{n-1} - \mu_{n-1}}{v_{n-1}} \right] \quad (8)$$

where $\bar{g}(\mu_n)$ reduces to μ_n and $Y_n^* = Y_n$.

Other models are contained in this general class, such as those in [8, 9, 11, 19].

3 Strict stationarity for the general model

In this section we present results for stationarity conditions of the chain $\{X_n\}_{n \in \mathbb{N}}$ and the process $\{Y_n\}_{n \in \mathbb{N}}$ coming from (5).

The usual practical condition for establishing stationarity and ergodicity in a Markov chain is by showing that it is positive Harris recurrent, via a drift condition in “small set”. Positive Harris recurrent chains possess a unique stationary probability distribution π . However, this does not work if the chain is not \emptyset -irreducible, as for the case of $\{Y_n\}_{n \in \mathbb{N}}$ integer-valued (for the details, see [14]). Nevertheless, as suggested by [13], one can still use the drift condition combined with the *weak Feller* property to show existence of a stationary distribution. Then, by applying the *asymptotic strong Feller* condition, one can derive uniqueness of the stationary distribution (for the definitions, see [18, 10, 13]).

Let $E_x(\cdot)$ denote the expectation under the probability $P_x(\cdot)$ induced on the path space of the chain when the initial state is $X_0 = x$.

We handle three separate cases:

1. $f(\cdot; \mu)$ is defined for any $\mu \in \mathbb{R}$. In this case the domain of g and h is \mathbb{R} and $Y_n^* = Y_n$ is taken;
2. $f(\cdot; \mu)$ is defined for only $\mu \in \mathbb{R}^+$ (or μ on any one-sided open interval by analogy). In this case the domain of g and h is \mathbb{R}^+ and $Y_n^* = \max\{Y_n, c\}$ for some $c \geq 0$ is taken;
3. $f(\cdot; \mu)$ is defined for only $\mu \in (0, a)$ where $a > 0$ (or any bounded interval by analogy). In this case the domain of g and h is $(0, a)$ and for some $c \in [0, a/2)$ $Y_t^* = \min\{\max(Y_n, c), (a - c)\}$ is taken.

Let $Y_0(x)$ denote the random variable Y_0 conditional on $\mu_0 = x$.

Definition 1. The Lipschitz condition

$$|\tilde{g}(z) - \tilde{g}(w)| \leq L|z - w| \quad (9)$$

with $L \leq 1$, is satisfied in the following different scenarios:

1. $\bar{g} \equiv h \neq g$
2. $\bar{g} \neq g$ and $h : \text{identity}$
3. $E[h(Y_t^*) | \mathcal{F}_{t-1}] = \bar{g}(\mu_t) \neq g(\mu_t)$

under the assumption that the link functions g^{-1} , h , \bar{g} are Lipschitz with constant smaller or equal than 1.

Theorem 1. *The process $\{\mu_n\}_{n \in \mathbb{N}}$ specified by the model (5) has a stationary distribution, and thus is stationary for an appropriate initial distribution for μ_0 (then, $\{Y_n\}_{n \in \mathbb{N}}$ is stationary), under the conditions below.*

1. $Y_0(x) \Rightarrow Y_0(x')$ as $x \rightarrow x'$.
2. $E(Y_n | \mu_n) = \mu_n$.
3. There exist $\delta > 0$, $r \in [0, 1 + \delta)$ and nonnegative constants d_1, d_2 such that

$$E(|Y_n - \mu_n|^{2+\delta} | \mu_n) \leq d_1 |\mu_n|^r + d_2.$$

4. g and h are bijective and increasing, and

- If $\bar{g}(\mu_t) = g(\mu_t)$,

Stationarity of general class of observation driven models for discrete valued processes

- a. $h : \mathbb{R} \mapsto \mathbb{R}$ concave on \mathbb{R}^+ and convex on \mathbb{R}^- , $g : \mathbb{R} \mapsto \mathbb{R}$ concave on \mathbb{R}^+ and convex on \mathbb{R}^- , and $|\phi| + |\gamma| < 1$
- b. $h : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on \mathbb{R}^+ , $g : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on \mathbb{R}^+ , and $(|\gamma| + |\phi|) \vee |\theta + \gamma| < 1$
- c. $|\theta + \gamma| < 1$; no additional conditions on $h : (0, a) \mapsto \mathbb{R}$ and $g : (0, a) \mapsto \mathbb{R}$.
- If $\tilde{g}(\mu_t) \neq g(\mu_t)$ and \tilde{g} satisfies the Lipschitz condition (9),
 - a. $h : \mathbb{R} \mapsto \mathbb{R}$ concave on \mathbb{R}^+ and convex on \mathbb{R}^- , $g : \mathbb{R} \mapsto \mathbb{R}$ concave on \mathbb{R}^+ and convex on \mathbb{R}^- , and $|\phi| + |\gamma| < 1$
 - b. $h : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on \mathbb{R}^+ , $g : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on \mathbb{R}^+ , and $|\gamma| + (|\phi| \vee |\theta|) < 1$
 - c. $|\theta| + |\gamma| < 1$; no additional conditions on $h : (0, a) \mapsto \mathbb{R}$ and $g : (0, a) \mapsto \mathbb{R}$.

Let $X_n = g(\mu_n)$. For $X_0 = x$ and $g(\mu) = x$ we have that

$$X_1(x) = \alpha + \phi h(Y_0^*(g^{-1}(x))) + \theta[h(Y_0^*(g^{-1}(x))) - \tilde{g}(x)] + \gamma x$$

where $\tilde{g}(x) = (\tilde{g} \circ g^{-1})(x) = \tilde{g}(g^{-1}(x)) = \tilde{g}(\mu)$.

Since g^{-1} is continuous, $Y_0(g^{-1}(x)) \Rightarrow Y_0(g^{-1}(x'))$ as $x \rightarrow x'$. Since the $*$ that maps Y_0 to the domain of h is continuous, it follows that $Y_0^*(g^{-1}(x)) \Rightarrow Y_0^*(g^{-1}(x'))$ as $x \rightarrow x'$. Since h is continuous, we have that $h(Y_0^*(g^{-1}(x))) \Rightarrow h(Y_0^*(g^{-1}(x')))$. Since $\tilde{g}(x)$ is continuous, we have that $\tilde{g}(x) \Rightarrow \tilde{g}(x')$. So $X_1(x) \Rightarrow X_1(x')$ as $x \rightarrow x'$, showing the weak Feller property. So by combining this fact and Theorem 1, Theorem in [18] is satisfied and, then, a stationary distribution for $\{\mu_n\}_{n \in \mathbb{N}}$ and $\{Y_n\}_{n \in \mathbb{N}}$ exists.

Assume that the distribution $\pi_z(\cdot)$ of $g(Y_n^*)$ conditional on $g(\mu_t) = z$ has the Lipschitz property

$$\sup_{w, z \in \mathbb{R}: w \neq z} \frac{\|\pi_w(\cdot) - \pi_z(\cdot)\|_{TV}}{|w - z|} < B < \infty \quad (10)$$

where $\|\cdot\|_{TV}$ is the total variation norm (see [14], pag. 315).

Theorem 2. Suppose that the conditions of Theorem 1 and the Lipschitz condition (10) hold, and that there is some $x \in \mathbb{R}$ that is in the support of Y_0 for all values of μ_0 . Then, there is a unique stationary distribution for $\{\mu_n\}_{n \in \mathbb{N}}$.

Proof. A sketch of the proofs of Theorems 1 and 2 is postponed to the Appendix.

4 Strict stationarity of specific models

In this section, the results obtained in Theorems 1 and 2 are applied to specific models of potential interest.

We first remark that for the GARMA model (6), Theorems 1 and 2 reduce exactly to the results of [13].

For the BARMA model, Theorem 1 and 2 reduce to the following proposition.

Proposition 1. *Suppose that conditional on μ_n , Y_n is Binomial(n, μ_n), with fixed number of trials n , the link function $g : (0, a) \mapsto \mathbb{R}$ is bijective and increasing, g^{-1} is Lipschitz and $|\theta| < 1$. Then, the process $\{\mu_n\}_{n \in \mathbb{N}}$ defined in (7) has a unique stationary distribution π . Hence, when μ_0 is initialized according to π , the process $\{Y_n\}_{n \in \mathbb{N}}$ is strictly stationary.*

In [13] (pag. 820-821) is proved that, for Poisson and Binomial distributions, the Lipschitz conditions (10) holds when g^{-1} is Lipschitz. We proved that the same holds for the Negative Binomial distribution. Note that the conditions on g and g^{-1} are clearly satisfied for the usual link, like logit or probit.

For GLARMA models, no stationarity results are available, apart from the simplest case when $k = 0$, $q = 1$ (see [3], [7], [4]). Our Theorems 1 and 2 imply the following proposition.

Proposition 2. *The process $\{\mu_n\}_{n \in \mathbb{N}}$ specified by the model (8), has a unique stationary distribution π , and thus is stationary when μ_0 is initialized according to π , under the conditions below. This implies that $\{Y_n\}_{n \in \mathbb{N}}$ is strictly stationary when μ_0 is initialized according to π . The conditions are:*

1. $E(Y_n | \mu_n) = \mu_n$.
2. $(2 + \delta \text{ moment condition})$: There exist $\delta > 0$, $r \in [0, 1 + \delta)$ and nonnegative constants d_1, d_2 such that

$$E(|Y_n - \mu_n|^{2+\delta} | \mu_n) \leq d_1 |\mu_n|^r + d_2.$$

3. g is bijective and increasing, and

- a. $g : \mathbb{R} \mapsto \mathbb{R}$ concave on \mathbb{R}^+ and convex on \mathbb{R}^- , and $|\gamma| < 1$
- b. $g : \mathbb{R}^+ \mapsto \mathbb{R}$ concave on \mathbb{R}^+ , and $|\gamma| + |\theta| < 1$
- c. $|\gamma| + |\theta| < 1$; no additional conditions on $g : (0, a) \mapsto \mathbb{R}$.

4. g^{-1} is Lipschitz with constant not greater than 1.
5. If $\{Y_t\}_{t \in \mathbb{N}}$ is discrete-valued, then (10) need to hold.

Note that, in the GLARMA model, the conditional distribution of $\{Y_t\}_{t \in \mathbb{N}}$ belongs to the exponential family, thus the first two moment conditions are satisfied. As mentioned above, for usual choices of discrete distributions (Poisson, Binomial, or Negative Binomial) the Lipschitz conditions (10) holds when g^{-1} is Lipschitz.

Finally, the conditions on g and g^{-1} clearly hold for the usual link functions.

In practical applications, one just needs to verify the condition on the coefficients to establish the stationarity of the model.

5 Concluding remarks and further developments

This paper provides a framework for proving the existence of stationarity and ergodic solutions for a wide class of observation-driven time series models. For many models in the class, no such results were available in the literature.

The Lipschitz assumption (10) is not satisfied when the y -link function h is the logarithmic function. In [5], different assumptions are considered to weaken the Lipschitz condition. We shall investigate them in the future.

All the models encompassed are often used with covariates. Extending our results to accomplish for covariates would be a further aspect to investigate.

Finally, our stationarity results could be used for proving consistency and asymptotic normality of estimators in discrete-valued models. The results of [5] and [6] could be used in future work with the aim to develop the asymptotic theory for the class of models considered in this paper.

Appendix

The proof for Theorem 1 and 2 follows the line of Theorems 5 and 15 in [13]. We provide a sketch the proof here in the following.

Having showed that the set $A = [-M; M]$, $M > 0$, is a small set, it is possible to prove a drift condition by taking the energy function $V(x) = |x|$ for the model (5):

$$\mathbb{E}_x V(X_1) = \mathbb{E}_x |\alpha + \gamma x + \phi h(Y_0^*) + \theta [h(Y_0^*) - \bar{g}(\mu)]| \quad (11)$$

and find that it is bounded under certain conditions on the coefficients, in the same fashion of Theorem 5 of [13]. For sake of brevity, we omit the details.

The last step required for completing the proof it to show that the Markov chain $\{X_t\}_{t \in \mathbb{N}}$ is asymptotically strong Feller. This is accomplished by a modification of the proof for Theorem 15 of [13]: set $g \equiv \bar{g}$, then, the random variables $g(Y_0^*(z))$ and $g(Y_0^*(w))$ have marginal distributions π_z and π_w , and $\mathbb{P}(g(Y_0^*(w)) = g(Y_0^*(z))) = 1 - \|\pi_w(\cdot) - \pi_z(\cdot)\|_{TV} > 1 - B|z - w|$.

If $h(Y_0^*(w)) = h(Y_0^*(z))$ then $|Z_1(w) - Z_1(z)| = |-\theta(\bar{g}(w) - \bar{g}(z)) + \gamma(z - w)| = |\theta + \gamma||z - w|$ and so $\|\pi_{Z_1(z)}(\cdot) - \pi_{Z_1(w)}(\cdot)\|_{TV} < B|Z_1(z) - Z_1(w)| < B|\theta + \gamma||z - w|$. Then we can construct $g(Y_1^*(z))$ and $g(Y_1^*(w))$ so that they have the correct marginal distributions and that $\mathbb{P}(g(Y_1^*(w)) = g(Y_1^*(z)) | g(Y_0^*(w)) = g(Y_0^*(z))) = \mathbb{P}(g(Y_1^*(w)) = g(Y_1^*(z)) | h(Y_0^*(w)) = h(Y_0^*(z))) > 1 - \|\pi_{Z_1(z)}(\cdot) - \pi_{Z_1(w)}(\cdot)\|_{TV} > 1 - B|\theta + \gamma||z - w|$ where the first equality works because g and h are one-to-one functions

$$g(Y_0^*(w)) = g(Y_0^*(z)) \iff Y_0^*(w) = Y_0^*(z) \iff h(Y_0^*(w)) = h(Y_0^*(z)).$$

If $h(Y_1^*(z)) = h(Y_1^*(w))$ then we can continue to “couple” the chains as above.

Notice that the probability that the chains couple for all times $0, 1, \dots$ is at least

$$1 - B|z - w| \sum_{t=0}^{\infty} (|\theta + \gamma|)^t = 1 - \frac{|z - w|B}{1 - |\theta + \gamma|},$$

where the inequality applies by imposing $|\theta + \gamma| < 1$. Thus, we combine this coefficient condition with those obtained from the drift condition (11) into Theorem 1. The remaining follows as in [13].

If $g \neq \bar{g}$ is it possible to adapt the previous proof in the following way
 $|Z_1(w) - Z_1(z)| = |-\theta(\bar{g}(w) - \bar{g}(z)) + \gamma(z - w)| \leq |\theta||\bar{g}(w) - \bar{g}(z)| + |\gamma||z - w|$ and,
under the Lipschitz condition (9) we obtain $|Z_1(w) - Z_1(z)| \leq |\theta||\bar{g}(w) - \bar{g}(z)| + |\gamma||z - w| \leq (|\theta| + |\gamma|)|z - w|$. Hence, the proof for the former case $\bar{g} \equiv g$ works
also for other shapes of \bar{g} , by substituting the condition $|\theta + \gamma|$ with $|\theta| + |\gamma|$ and
by combining it in Theorem 1. Clearly, the condition (9) depends on the shape of
 \bar{g} . However, it is easy to show that it works for all the specific shapes of the link
functions considered in Definition 1. We omit the details.

References

1. Benjamin, M.A., Rigby, R.A., Stasinopoulos, D.M.: Generalized autoregressive moving average models. *J. Amer. Stat. Assoc.* **98**, (461), 214–223 (2003)
2. Cox, D.: Statistical analysis of time-series: some recent developments. *Scand. J. Stat.* **8**, (2), 93–115 (1981)
3. Davis, R., Dunsmuir, W., Streett, S.: Observation-driven models for Poisson counts. *Biometrika* **90**, (4), 777–790 (2003)
4. Davis, R., Liu, H.: Theory and inference for a class of observation-driven models with application to time series of counts. *Stat. Sin.* **26**, (4), 1673–1707 (2016)
5. Douc, R., Doukhan, P., Moulines, E.: Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stoch. Proc. Appl.* **123**, (7), 2620–2647 (2013)
6. Douc, R., Fokianos, K., Moulines, E.: Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models. *Electron. J. Stat.* **11**, (2), 2707–2740 (2017)
7. Dunsmuir, W.T., Scott, D.J.: The glarma package for observation driven time series regression of counts. *J. Stat. Soft.* **67**, (7), 1–36 (2015)
8. Ferland, R., Latour, A., Oraichi, D.: Integer-valued GARCH process. *J. Time Ser. Anal.* **27**, (6), 923–942 (2006)
9. Fokianos, K., Tjøstheim, D.: Log-linear Poisson autoregression. *J. Multivar. Anal.* **102**, (3), 563–578 (2011)
10. Hairer, M., Mattingly, J.C.: Ergodicity of the 2D Navier-Stokes equations with degenerate stochastic forcing. *Ann. Math.* **164** 993–1032 (2006)
11. Kauppi, H., Saikkonen, P.: Predicting US recessions with dynamic binary response models. *Rev. Econ. Stat.* **90**, (4), 777–791 (2008)
12. Li, W.K.: Time series models based on generalized linear models: some further results. *Biometrics* **50**, (2), 506–511 (1994)
13. Matteson, D.S., Woodard, D.B., Henderson, S.G.: Stationarity of generalized autoregressive moving average models. *Electron. J. Stat.* **5**, 800–828 (2011)
14. Meyn, S., Tweedie, R.L., Glynn, P. W.: *Markov Chains and Stochastic Stability*. Cambridge University Press, London (2009)
15. Roberts, G.O., Rosenthal, J.S.: General state space markov chains and mcmc algorithms. *Probab. Surv.* **1**, 20–71 (2004)
16. Rydberg, T.H., Shephard N.: Dynamics of trade-by-trade price movements: Decomposition and models. *J. Financ. Economet.* **1**, (1), 2–25 (2003)
17. Startz, R.: Binomial autoregressive moving average models with an application to U.S. recessions. *J. Bus. Econ. Stat.* **26**, (1), 1–8 (2008)
18. Tweedie, R.L.: Invariant measures for Markov chains with no irreducibility assumptions. *J. Appl. Probab.* **25**, (A), 275–285 (1988)
19. Zheng, T., Xiao, H., Chen, R.: Generalized ARMA models with martingale difference errors. *J. Economet.* **189**, (2), 492–506 (2015)

An extension of the censored gaussian lasso estimator

Un'estensione dello stimatore clgasso

Luigi Augugliaro and Gianluca Sottile and Veronica Vinciotti

Abstract The conditional glasso is one of the most used estimators for inferring genetic networks. Despite its diffusion, there are several fields in applied research where the limits of detection of modern measurement technologies make the use of this estimator theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. In this paper we propose an extension to censored data.

Abstract *Il conditional graphical lasso è uno degli stimatori più utilizzati per fare inferenza sulle reti genetiche. Nonostante la sua elevata diffusione, esistono parecchi campi applicativi dove i limiti degli strumenti di misurazione ne rendono teoricamente ingiustificato l'utilizzo, anche quando l'assunzione relativa alla distribuzione normale multivariata è soddisfatta.*

Key words: Censored data, Censored glasso estimator, Gaussian graphical model, glasso estimator.

1 Introduction

An important aim in genomics is to understand interactions among genes, characterized by the regulation and synthesis of proteins under internal and external signals. These relationships can be represented by a genetic network, i.e., a graph where

Luigi Augugliaro

Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: luigi.augugliaro@unipa.it

Gianluca Sottile

Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: gianluca.sottile@unipa.it

Veronica Vinciotti

Department of Mathematics, Brunel University, UK, e-mail: veronica.vinciotti@brunel.ac.uk

nodes represent genes and edges describe the interactions among them. Gaussian graphical models [3] have been widely used for reconstructing a genetic network from expression data. The reason of such diffusion relies on the statistical properties of the multivariate Gaussian distribution which allow the topological structure of a network to be related with the non-zero elements of the concentration matrix, i.e., the inverse of the covariance matrix. Thus, the problem of network inference can be recast as the problem of estimating a concentration matrix. The conditional glasso estimator [8] is a popular method for estimating a sparse concentration matrix, based on the idea of adding an ℓ_1 -penalty function to the likelihood function of the multivariate Gaussian distribution.

Despite the widespread literature on the conditional glasso estimator, there is a great number of fields in applied research where modern measurement technologies make the use of this graphical model theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. A first example of this is Reverse Transcription quantitative Polymerase Chain Reaction (RT-qPCR), a popular technology for gene expression profiling. This technique relies on fluorescence-based detection of amplicon DNA and allows the kinetics of PCR amplification to be monitored in real time, making it possible to quantify nucleic acids with extraordinary ease and precision. The analysis of the raw RT-qPCR profiles is based on the cycle-threshold, defined as the fractional cycle number in the log-linear region of PCR amplification in which the reaction reaches fixed amounts of amplicon DNA. If a target is not expressed or the amplification step fails, the threshold is not reached after the maximum number of cycles (limit of detection) and the corresponding cycle-threshold is undetermined. For this reason, the resulting data is naturally right-censored data. In this paper we propose an extension of the conditional glasso estimator that takes into account the censoring mechanism of the data explicitly.

2 The conditional censored Gaussian graphical model

Let $\mathbf{Y} = (Y_1, \dots, Y_p)^\top$ be a p -dimensional random vector. Graphical models allow to represent the set of conditional independencies among these random variables by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where \mathcal{V} is the set of nodes associated to \mathbf{Y} and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of ordered pairs, called edges, representing the conditional dependencies among the p random variables [3]. The conditional Gaussian graphical model is an extension of the classical Gaussian graphical model based on the assumption that the conditional distribution of \mathbf{Y} given a q -dimensional vector of predictors, say $\mathbf{X} = (X_1, \dots, X_q)^\top$, follows a multivariate Gaussian distribution with expected value:

$$\boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \mathbf{x},$$

where $\boldsymbol{\beta} = (\beta_{hk})$ is a matrix $q \times p$ coefficient matrix, and covariance matrix denoted by $\boldsymbol{\Sigma} = (\sigma_{hk})$. Denoting with $\boldsymbol{\Theta} = (\theta_{hk})$ the concentration matrix, i.e., the inverse

of the covariance matrix, the conditional density function of \mathbf{Y} can be written as follows:

$$\phi(\mathbf{y} | \mathbf{x}; \boldsymbol{\beta}, \Theta) = (2\pi)^{-p/2} |\Theta|^{1/2} \exp[-1/2 \{\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}^\top \Theta \{\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}]. \quad (1)$$

As shown in [3], the off-diagonal elements of the concentration matrix are the parametric tools relating the pairwise Markov property to the factorization of the density (1). Formally, two random variables, say Y_h and Y_k , are conditionally independent given all the remaining variables if and only if θ_{hk} is equal to zero. This result provides a simple way to relate the topological structure of the graph \mathcal{G} to the pairwise Markov property, i.e., the undirected edge (h, k) is an element of the edge set \mathcal{E} if and only if $\theta_{hk} \neq 0$,

As done in [1], we assume that \mathbf{Y} is a (partially) latent random vector with density function (1). In order to include the censoring mechanism inside our framework, let us denote by $\mathbf{l} = (l_1, \dots, l_p)^\top$ and $\mathbf{u} = (u_1, \dots, u_p)^\top$, with $l_h < u_h$ for $h = 1, \dots, p$, the vectors of known left and right censoring values. Thus, Y_h is observed only if it is inside the interval $[l_h, u_h]$ otherwise it is censored from below if $Y_h < l_h$ or censored from above if $Y_h > u_h$. Under this setting, a rigorous definition of the joint distribution of the observed data can be obtained using the approach for missing data with nonignorable mechanism [4]. This requires the specification of the distribution of a p -dimensional random vector, denoted by $R(\mathbf{Y}; \mathbf{l}, \mathbf{u})$, used to encode the censoring patterns. Formally, the h th element of $R(\mathbf{Y}; \mathbf{l}, \mathbf{u})$ is defined as $R(Y_h; l_h, u_h) = I(Y_h > u_h) - I(Y_h < l_h)$, where $I(\cdot)$ denotes the indicator function. By construction $R(\mathbf{Y}; \mathbf{l}, \mathbf{u})$ is a discrete random vector with support the set $\{-1, 0, 1\}^p$ and probability function $\Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r}\} = \int_{D_{\mathbf{r}}} \phi(\mathbf{y} | \mathbf{x}; \boldsymbol{\beta}, \Theta) d\mathbf{y}$, where $D_{\mathbf{r}} = \{\mathbf{y} \in \mathbb{R}^p : R(\mathbf{y}; \mathbf{l}, \mathbf{u}) = \mathbf{r}\}$.

Given a censoring pattern, we can simplify our notation by partitioning the set $\mathcal{J} = \{1, \dots, p\}$ into $o = \{h \in \mathcal{J} : r_h = 0\}$, $c^- = \{h \in \mathcal{J} : r_h = -1\}$ and $c^+ = \{h \in \mathcal{J} : r_h = +1\}$ and, in the following of this paper, we shall use the convention that a vector indexed by a set of indices denotes the corresponding subvector. For example, the subvector of observed elements in \mathbf{y} is denoted by $\mathbf{y}_o = (y_h)_{h \in o}$ and, consequently, the observed data is the vector $(\mathbf{y}_o^\top, \mathbf{x}^\top, \mathbf{r}^\top)^\top$. As done in [1], the probability distribution of the observed data, denoted by $\varphi(\{\mathbf{y}_o, \mathbf{r}\} | \mathbf{x}; \boldsymbol{\beta}, \Theta)$, can be defined as follows:

$$\varphi(\{\mathbf{y}_o, \mathbf{r}\} | \mathbf{x}; \boldsymbol{\beta}, \Theta) = \int \phi(\{\mathbf{y}_o, \mathbf{y}_c\} | \mathbf{x}; \boldsymbol{\beta}, \Theta) \Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r} | \mathbf{Y} = \mathbf{y}\} d\mathbf{y}_c, \quad (2)$$

where $c = c^- \cup c^+$.

Density (2) can be simplified by observing that $\Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r} | \mathbf{Y} = \mathbf{y}\}$ is equal to one if the censoring pattern encoded in \mathbf{r} is equal to the pattern observed in \mathbf{y} , otherwise it is equal to zero, i.e.,

$$\Pr\{R(\mathbf{Y}; \mathbf{l}, \mathbf{u}) = \mathbf{r} | \mathbf{Y} = \mathbf{y}\} = I(\mathbf{y}_{c^-} < \mathbf{l}_{c^-}) I(\mathbf{l}_o \leq \mathbf{y}_o \leq \mathbf{u}_o) I(\mathbf{u}_{c^+} < \mathbf{y}_{c^+}),$$

where the inequalities in the previous expressions are intended elementwise. From this, $\varphi(\{\mathbf{y}_o, \mathbf{r}\} | \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\Theta})$ can be rewritten as

$$\varphi(\{\mathbf{y}_o, \mathbf{r}\} | \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\Theta}) = \int_{D_c} \phi(\{\mathbf{y}_o, \mathbf{y}_c\} | \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\Theta}) d\mathbf{y}_c I(\mathbf{l}_o \leq \mathbf{y}_o \leq \mathbf{u}_o), \quad (3)$$

where $D_c = (-\infty, \mathbf{l}_{c^-}) \times (\mathbf{u}_{c^+}, +\infty)$. Using density (3), the conditional censored Gaussian graphical model is defined as the set $\{\mathbf{Y}, R(\mathbf{Y}; \mathbf{l}, \mathbf{u}), \varphi(\{\mathbf{y}_o, \mathbf{r}\} | \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\Theta}), \mathcal{G}\}$, where $\varphi(\{\mathbf{y}_o, \mathbf{r}\} | \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\Theta})$ factorizes according to the undirected graph \mathcal{G} .

3 The conditional censored glasso estimator

Suppose we have a sample of size n independent observations drawn from a conditional censored Gaussian graphical model. For ease of exposition, we shall assume that \mathbf{l} and \mathbf{u} are fixed across the n observations, but the extension to the cases where the censoring vectors are specific to each observation is straightforward and does not require a specific treatment. To simplify our notation the set of indices of the variables observed in the i th observation is denoted by $o_i = \{h \in \mathcal{J} : r_{ih} = 0\}$, while $c_i^- = \{h \in \mathcal{J} : r_{ih} = -1\}$ and $c_i^+ = \{h \in \mathcal{J} : r_{ih} = +1\}$ denote the sets of indices associated to the left and right-censored data, respectively. Denoting by \mathbf{r}_i the realization of the random vector $R(\mathbf{Y}_i; \mathbf{l}, \mathbf{u})$, the i th observed data is the vector $(\mathbf{y}_{io_i}^\top, \mathbf{x}_i^\top, \mathbf{r}_i^\top)^\top$. Using the density function (3), the observed log-likelihood function can be written as

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \sum_{i=1}^n \log \int_{D_{c_i}} \phi(\{\mathbf{y}_{io_i}, \mathbf{y}_{ic_i}\} | \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\Theta}) d\mathbf{y}_{ic_i} = \sum_{i=1}^n \log \varphi(\{\mathbf{y}_{io_i}, \mathbf{r}_i\} | \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\Theta}), \quad (4)$$

where $D_{c_i} = (-\infty, \mathbf{l}_{c_i^-}) \times (\mathbf{u}_{c_i^+}, +\infty)$ and $c_i = c_i^- \cup c_i^+$. Although inference about the parameters of this model can be carried out via the maximum likelihood method, the application of this inferential procedure to real datasets is limited for three main reasons. Firstly, the number of measured variables is often larger than the sample size and this implies the non-existence of the maximum likelihood estimator even when the dataset is fully observed. Secondly, even when the sample size is large enough, the maximum likelihood estimator will exhibit a very high variance [5, 7]. Thirdly, empirical evidence suggests that gene networks or more general biochemical networks are not fully connected [2]. In terms of conditional Gaussian graphical models this evidence translates in the assumption that $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ have a sparse structure, i.e., only few regression coefficients and few θ_{hk} are different from zero.

All that considered, we propose to estimate the parameters of the conditional censored Gaussian graphical model by generalizing the approach proposed in [8], i.e., by maximizing a new objective function defined by adding two lasso-type penalty functions to the observed log-likelihood (4). The resulting estimator, called conditional censored glasso estimator, is formally defined as

$$\{\hat{\boldsymbol{\beta}}^\lambda, \hat{\boldsymbol{\Theta}}^\rho\} = \arg \max_{\boldsymbol{\beta}, \boldsymbol{\Theta} \succ 0} \frac{1}{n} \sum_{i=1}^n \log \varphi(\{\mathbf{y}_{i_{oi}}, \mathbf{r}_i\} | \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\Theta}) - \lambda \sum_{h,k} |\beta_{hk}| - \rho \sum_{h \neq k} |\theta_{hk}|, \quad (5)$$

where λ and ρ are two non-negative tuning parameters. The lasso penalty on $\boldsymbol{\beta}$ introduces sparsity in $\hat{\boldsymbol{\beta}}^\lambda$, in other words by varying λ we can select the relevant predictors for \mathbf{Y} . Like in the standard glasso estimator, the tuning parameter ρ controls the amount of sparsity in the estimated concentration matrix $\hat{\boldsymbol{\Theta}}^\rho = (\hat{\theta}_{hk}^\rho)$ and, consequently, in the corresponding estimated graph $\hat{\mathcal{G}}^\rho = \{\mathcal{V}, \hat{\mathcal{E}}^\rho\}$, where $\hat{\mathcal{E}}^\rho = \{(h, k) : \hat{\theta}_{hk}^\rho \neq 0\}$. When ρ is large enough, some $\hat{\theta}_{hk}^\rho$ are shrunk to zero resulting in the removal of the corresponding link in $\hat{\mathcal{G}}^\rho$; on the other hand, when ρ is equal to zero and the sample size is large enough the estimator $\hat{\boldsymbol{\Theta}}^\rho$ coincides with the maximum likelihood estimator of the concentration matrix, which implies a fully connected estimated concentration graph.

4 Simulation study

In this section, we compare our proposed estimator with MissGlasso [6], which performs ℓ_1 -penalized estimation under the assumption that the censored data are missing at random, and with the conditional glasso estimator [8], where the empirical covariance matrix is calculated by imputing the missing values with the censoring values. These estimators are evaluated in terms of both recovering the structure of the true graph and the mean squared error. We use the method implemented in the R package `huge` [9], to simulate a sparse concentration matrix with a random structure for \mathbf{Y} . In particular, we set the probability of observing a link between two nodes to k/p , where p is the number of responses and k is used to control the amount of sparsity in $\boldsymbol{\Theta}$. Moreover, we set the right censoring value to 40 for any variable and the sample size n to 100. The predictors matrix \mathbf{X} is sampled from a multivariate gaussian distribution with zero expected value and sparse covariance matrix simulated as done for \mathbf{Y} . Each column the true matrix of predictors $\boldsymbol{\beta}$ contains only two non-zero regression coefficients, the values are sampled from a uniform distribution on the interval $[0.3, 0.7]$. The values of the intercepts are chosen in such a way that H response variables are right censored with probability equal to 0.40. The quantities k, p, q and H are used to specify the different scenarios used to analyze the behavior of the considered estimators. In particular, we consider the following cases:

- **Scenario 1:** $k = 3, p = 50, q = 10$ and $H = 25$. This setting is used to evaluate the effects of the number of censored variables on the behavior of the proposed estimators when $n > p$.
- **Scenario 2:** $k = 3, p = 150, q = 10$ and $H = 75$. This setting is used to evaluate the impact of the high dimensionality on the estimators ($p > n$).

For each scenario, we simulate 50 samples and in each simulation, we compute the coefficients path using `cglasso`, `MissGlasso`, and `glasso`. Each path is computed

using an equally spaced sequence of ρ and λ -values. However, the two scenarios differ also on the length of the two sequences, that is 20 for the Scenario 1 and 10 for the Scenario 2. Moreover, the precision-recall curves and the area under the curves (AUCs) are computed for each Scenarios.

The curves report the relationship between precision and recall for any ρ and λ -value, which are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where TP, FP, and FN are quantities defined as the number correctly selected non-null items, the number of wrong selected non-null items and the number of wrong selected null item, respectively. Table 1 shows how cglasso gives a better estimate of the concentration and coefficient matrices in terms of AUCs, for any given value of the tuning parameters. We report only five evenly spaced values of λ and ρ . Figures 1 and 2 show the averages of the quantities $\min_{\rho} \text{MSE}(\hat{\Theta}_{\rho})$ and $\min_{\lambda} \text{MSE}(\hat{\beta}_{\lambda})$, which gives the minimum value of the mean squared error attained along the path of solutions. These plots emphasize that cglasso has also a mean squared error much smaller than the considered competitors.

Table 1 Mean area under the curves across the sequence of ρ and λ -values under the specification of the two Scenarios. The first column block refers to the concentration matrix (Θ) when λ is fixed and the second refers to the coefficient matrix (β) when ρ is fixed.

		λ/λ_{\max}					ρ/ρ_{\max}				
		0.00	0.25	0.50	0.75	1.00	0.00	0.25	0.50	0.75	1.00
Model 1	cglasso	0.546	0.429	0.139	0.103	0.101	0.844	0.877	0.883	0.882	0.885
	MissGlasso	0.239	0.199	0.086	0.073	0.073	0.745	0.764	0.766	0.767	0.768
	glasso	0.414	0.218	0.097	0.092	0.091	0.813	0.847	0.864	0.866	0.866
Model 2	cglasso	0.418	0.094	0.037	0.035	0.035	0.794	0.930	0.931	0.929	0.933
	MissGlasso	0.329	0.098	0.033	0.031	0.030	0.753	0.830	0.831	0.830	0.831
	glasso	0.321	0.040	0.033	0.032	0.031	0.751	0.902	0.906	0.907	0.907

5 Conclusions

In this paper, we have proposed an extension of the conditional glasso estimator to multivariate censored data. A simulation study showed that the proposed estimator overcomes the existing estimators both in terms of parameter estimation and of network recovery.

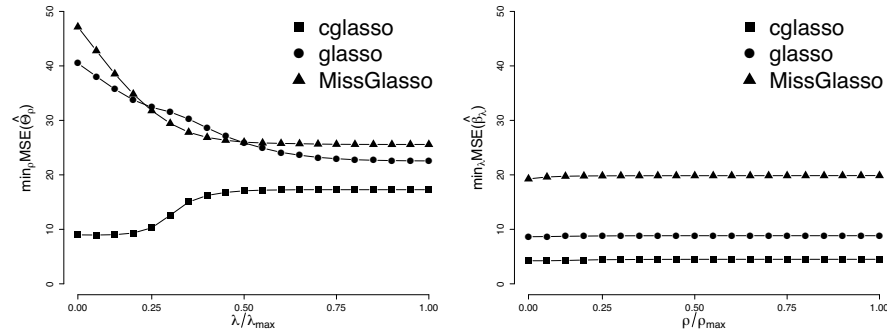


Fig. 1 Average of the minimum mean squared error attained along the path of solutions under the specification of Scenario 1. Left panel refers to the precision matrix for each fixed value of the tuning parameter λ ; right panel refers to the coefficient matrix for each fixed value of the tuning parameter ρ . The square refers to the cglasso estimator, the circle to the glasso estimator and the triangle to MissGlasso estimator.

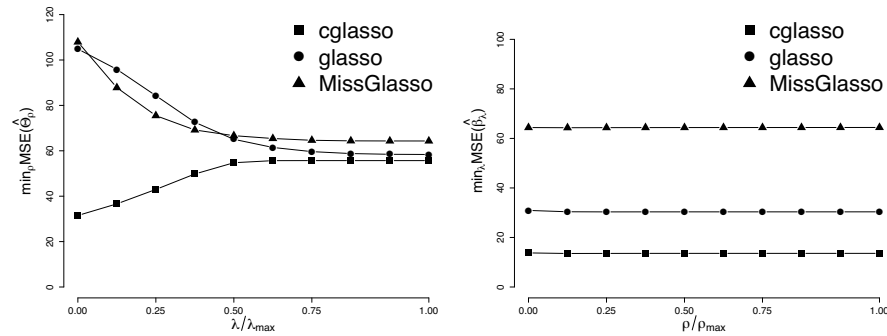


Fig. 2 Average of the minimum mean squared error attained, along the path of solutions under the specification of Model 2.

References

1. Augugliaro L., Abbruzzo A., Vinciotti V.: ℓ_1 -Penalized censored Gaussian graphical model. *Biostatistics*. <https://doi.org/10.1093/biostatistics/kxy043>
2. Gardner T. S., di Bernardo D., Lorenz D., Collins J. J.: Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*. **301**, 102–105 (2003)
3. Lauritzen S. L.: *Graphical Models*. Oxford University Press, Oxford (1996)
4. Little R. J. A., Rubin D. B.: *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., Hoboken (2002)
5. Schäfer J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*. **4**(1). (2005)
6. Städler, N., Bühlmann, P.: Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Stat. Comput.* **22**(1), 219–235 (2012)

7. Uhler C.: Geometry of maximum likelihood estimation in Gaussian graphical models. *Ann. Statist.* **40**(12), 238–261 (2012)
8. Yin J., Li H.: A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Statist.* **5**(4), 2630–2650 (2011)
9. Zhao T., Li X., Liu H., Roeder K., Lafferty J., Wasserman L.: huge: High-Dimensional Undirected Graph Estimation. R package version 1.2.7 (2015). <https://CRAN.R-project.org/package=huge>

A formal approach to data swapping and disclosure limitation techniques

Un approccio formale per tecniche di trasformazione dei dati in problemi di privacy

F. Ayed, M. Battiston and F. Camerlenghi

Abstract Data swapping is among the most popular disclosure limitation techniques for categorical data. Given a stochastic matrix M and a specified categorical variable, an individual belonging to category i is swapped to category j with probability $M_{i,j}$. In this work, we propose a formal way to choose the matrix M in order to balance between two desiderata: 1) preserving as much statistical information from the raw data as possible; 2) guaranteeing the privacy of individuals in the dataset. We propose to choose M as the solution of a constrained maximization problems, where we maximize the Mutual Information between raw and transformed data, given the constraint that the transformation satisfies the notion of Differential Privacy.

Abstract *Data swapping è tra gli approcci più popolari per la tutela della privacy in problemi con variabili categoriche. Data una matrice stocastica M e una specificata variabile categorica, il valore per un individuo appartenente alla categoria i viene cambiato in j con probabilità $M_{i,j}$. In questo lavoro proponiamo un metodo formale per scegliere M in modo che siano soddisfatti due requisiti: 1) preservare l'informazione statistica dei dati iniziali; 2) garantire la privacy degli individui nel dataset. Proponiamo di scegliere M come soluzione di un problema di massimizzazione vincolata, dove massimizziamo la Mutual Information tra il dataset iniziale e trasformato, sotto il vincolo che la trasformazione soddisfi la nozione di Differential Privacy.*

Key words: Data swapping, Disclosure risk, Mutual Information, Differential Privacy, Multinomial Models.

Fadhel Ayed
Oxford University, 24-29 St Giles', Oxford OX1 3LB, UK. e-mail: fadhel.ayed@gmail.com

Marco Battiston
Lancaster University, Fylde College, Bailrigg, Lancaster, LA1 4YF. UK e-mail: m.battiston@lancaster.ac.uk

Federico Camerlenghi
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy. e-mail: federico.camerlenghi@unimib.it

1 Introduction

When statistical and governmental agencies release microdata to the public, they often encounter ethical and moral issues concerning the possible privacy leak for individuals present in the dataset. Specifically, even after removing directly identifying variables, like names or national insurance numbers, an intruder might still be able to identify individuals by cross-classifying categorical variables in the dataset and matching them with some external database. This kind of privacy problems have been widely considered in the statistical literature and different measures of disclosure risk have been proposed to assess the riskiness of specific dataset. See [9, 10, 11, 2] for some references on disclosure risk estimations and [7] for a quite recent review.

Data swapping techniques are among the most used techniques for disclosure risk limitation. With these techniques, before releasing the dataset, the data curator swaps the values of some categorical identifying variables, like gender, job or age. In [9], the authors consider random swapping of a geographical variable and propose some measures of risk to assess whether random swapping has been effective in “privatizing” the dataset. The choice of the geographical variable is motivated by the fact that, by swapping it, it is usually less likely to generate unreasonable combinations of categorical variables, like for instance a pregnant man or a 10 year old lawyer. In order to implement data swapping, they introduce a stochastic matrix M , where the (i, j) entry of this matrix gives the probability that an individual from location i has his geographical variable swapped to location j . Given this known matrix M , Shlomo and Skinner (2010) [9] suggest some measures of risk and related estimation methods. However, it is unclear how the data curator should choose the matrix M in order to guarantee an effective level of privacy.

In this work we propose a formal way of choosing the swapping matrix M . Specifically, when choosing M , we need to balance two conflicting goals: 1) on the one hand, we want that the application of M to make the dataset somehow private; 2) on the other hand, we also want that the released dataset preserves as much statistical information as possible from the raw data. In order to balance this trade-off, we propose to choose M as the solution of a constrained maximization problem. We maximize the Mutual Information between the released and the raw dataset, hence guaranteeing preservation of statistical information and achieving goal 2). Mutual Information is a common measure of dependence between random variables used in probability and information theory, see Subsection 2.1. In order to guarantee also goal 1), we introduce a constrain in the maximization problem by imposing that the application M satisfies differential privacy. This notion provides a mathematical definition of privacy which has become quite popular in computer science over the last 10 years, see Subsection 2.2.

In Section 2, we describe the setting of the problem and briefly recall the definitions of mutual information and differential privacy. In Section 3, we introduce the proposed constrained maximization problem to choose M . First, in Subsection 3.1 we consider the case of swapping a binary variable, like gender or married/not

married variable. Then in Subsection 3.2, we start formalizing the general case of a categorical variable with n_k possible outcomes.

2 Disclosure risk, Mutual Information and Differential Privacy

In disclosure risk problems, we usually have microdata of n individuals, where for each individual we can observe two distinct kinds of variables: 1) some private variables, usually called *sensitive variables*, like his health status or salary; 2) some identifying categorical variables, usually called *key variables*, like gender, age, job. Disclosure problems arise because, by relying on external information, an intruder can identify individuals in a dataset by cross-classifying their key variables and therefore discover their sensitive information.

We assume to have J categorical key variables observed for a sample of n individuals. Each variable has n_j possible categories labelled from 1 up to n_j . The observation of individual i , $X_i = (X_{i1}, \dots, X_{iJ})$, takes values in the state space $\mathcal{C} := \prod_{j=1}^J \{1 \dots, n_j\}$. This set has $K := |\mathcal{C}| = \prod_j n_j$ values corresponding to all possible cross-classification of the J key variables. The information about the sample is usually given through the sample frequency vector (f_1, \dots, f_K) , where f_i counts how many individuals have been observed with that particular combination of cross-classified key variables.

When modelling (f_1, \dots, f_K) , the two most common choices are the *Poisson* and *Multinomial models* (see [1]). In Poisson models, the sample frequencies are assumed to be independent Poisson random variables

$$f_k \sim \text{Po}(\lambda_k).$$

In particular, in Poisson log-linear models, the parameters λ_k are assumed to depend on the particular values of the key variables through the formula $\lambda_k = \exp(\mathbf{x}_k \beta)$, for each $\mathbf{x}_k \in \mathcal{C}$, where β is a common unknown vector of J parameters to be estimated. In Multinomial models, we assume that the sample size is known and fixed. After conditioning to their sum, the Poisson frequency vector becomes Multinomial. Specifically, let $f_k \sim \text{Po}(\lambda_k)$ and $n = \sum_{i=1}^K f_i \sim \text{Po}(\sum_k \lambda_k)$, then $(f_1, \dots, f_K) | n \sim \text{Mult}(n, p_1, \dots, p_K)$, where $p_i = \frac{\lambda_i}{\sum_k \lambda_k}$.

2.1 Mutual Information

The *mutual information* between two discrete random variables X and Z is defined as

$$I(X, Z) = \sum_{z \in \mathcal{Z}} \sum_{x \in \mathcal{X}} P_{(X, Z)}(x, z) \log \left(\frac{P_{(X, Z)}(x, z)}{P_X(x) P_Z(z)} \right) \quad (1)$$

where \mathcal{X} , \mathcal{Z} and p_X, p_Z are respectively the state spaces and the marginal distributions of X and Z and $P_{(X,Z)}$ is their joint distribution on $\mathcal{X} \times \mathcal{Z}$. From the definition of $I(X, Z)$ it follows that

$$I(X, Z) = D_{KL}(p_{(X,Z)} || p_X p_Z) \quad (2)$$

where D_{KL} denotes the Kullback-Leibler divergence. Therefore, $I(X, Z)$ measures the divergence between the joint distribution of X and Z and the product of their marginals. From (2), it follows that $I(X, Z) \geq 0$, and $I(X, Z) = 0$ if and only if X and Z are independent. For a review on the mutual information and its properties, see for example [5, 6] and references therein.

2.2 Differential Privacy

Differential Privacy is a notion recently proposed in computer science [3, 4] to formalize the idea of private mechanism. See also [8] for a statistical viewpoint of differentail privacy.

Informally, let $X = (X_1, \dots, X_n)$ be a dataset containing sensitive information. A *mechanism* is simply a conditional distribution Q that, given the raw dataset X , returns a transformed dataset $Z = (Z_1, \dots, Z_{k_n})$ to be released to the public. If the mechanism is chosen carefully, it should be difficult for an intruder to recover the sensitive information inside X by only looking at Z .

Formally, we say that the conditional distribution Q of Z given X satisfies α -*Differential Privacy* if

$$\sup_{S \in \sigma(\mathcal{Z}^{k_n})} \frac{Q(Z_{1:k_n} \in S | X_{1:n})}{Q(Z_{1:k_n} \in S | X'_{1:n})} \leq \exp(\alpha). \quad (3)$$

for all $X_{1:n} = (X_1, \dots, X_n)$ and $X'_{1:n} = (X'_1, \dots, X'_n)$ s.t. $H(X_{1:n}, X'_{1:n}) = 1$, where H is the Hamming distance, $H(X_{1:n}, X'_{1:n}) = \sum_{i=1}^n I(X_i \neq X'_i)$.

3 An information-theoretic approach to data swapping

In [9] the authors consider the problem of disclosure risk estimation when the microdata has gone through a data swapping process. They swap the geographical key variable using a stochastic matrix M , i.e. every row of M sums to one, where M_{ij} provides the probability that an individual from location i is swapped to location j . The authors of [9] also discuss the problem of estimating some measures of risk, but without providing any tangible rule on how to choose M .

In this work, we propose to choose M as the solution of the following maximization problem:

A formal approach to data swapping and disclosure limitation techniques

$$\max_{M: Q \text{ satisfying (3)}} I(X_{1:n}, Z_{1:n}), \quad (4)$$

namely we look for M such that the information of $Z_{1:n}$ is as close as possible to that in $X_{1:n}$, and M is also constrained to satisfy differential privacy.

3.1 Binary key variable

We start from the simplest case of a categorical variable with only two possible categories denoted $\{0, 1\}$. Let $X = (X_1, \dots, X_n) \in \{0, 1\}^n$ denote the realization of the specific binary key variable for the n individuals in the microdata. We assume that X_i are sampled from a Multinomial model, hence $X_i \stackrel{iid}{\sim} \text{Bern}(p)$.

Let us now suppose to swap the value of each X_i independently with probability q . This means that we transform X into $Z = (Z_1, \dots, Z_n)$ through the conditional distribution

$$Z|X \sim Q(Z_{1:n}|X_{1:n}) = \prod_{i=1}^n q^{\mathbb{I}(Z_i=X_i)} (1-q)^{\mathbb{I}(Z_i \neq X_i)}$$

The corresponding swapping matrix M is therefore $[q, 1-q; 1-q, q]$. We are now going to find q , hence M , by solving (4).

The Marginal distributions of $X_{1:n}$ and $Z_{1:n}$ are respectively,

$$P(X_{1:n}; p) = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i},$$

$$P(Z_{1:n}; q, p) = \prod_{i=1}^n (pq + (1-q)(1-p))^{Z_i} ((1-q)p + q(1-p))^{1-Z_i}.$$

While the joint distribution is

$$P(X_{1:n}, Z_{1:n}; p, q) = Q(Z_{1:n}|X_{1:n}; q) P(X_{1:n}; p) = \prod_{i=1}^n q^{\mathbb{I}(Z_i=X_i)} (1-q)^{\mathbb{I}(Z_i \neq X_i)} p^{X_i} (1-p)^{1-X_i}$$

Considering $X_{1:n}$ and $X'_{1:n}$ s.t. $H(X_{1:n}, X'_{1:n}) = 1$ implies that at some coordinate i $X_i = 1$ and $X'_i = 0$ or viceversa, while all other coordinates are the same. Therefore condition (3) specializes as

$$\begin{aligned}
\sup_{S \in \sigma(\mathcal{Z}^n)} \frac{Q(Z_{1:n} \in S | X_{1:n})}{Q(Z_{1:n} \in S | X'_{1:n})} &= \sup_{S \subseteq \{0,1\}^n} \sum_{Z_{1:n} \in S} \frac{\prod_{i=1}^n q^{\mathbb{I}(Z_i=X_i)} (1-q)^{\mathbb{I}(Z_i \neq X_i)}}{\prod_{i=1}^n q^{\mathbb{I}(Z_i=X'_i)} (1-q)^{\mathbb{I}(Z_i \neq X'_i)}} \\
&= \sup_{S \subseteq \{0,1\}^n} \sum_{Z_{1:n} \in S} \frac{q^{\mathbb{I}(Z_i=X_i)} (1-q)^{\mathbb{I}(Z_i \neq X_i)}}{q^{\mathbb{I}(Z_i=X'_i)} (1-q)^{\mathbb{I}(Z_i \neq X'_i)}} \\
&= \sup_{Z_i \in \{0,1\}} \frac{q^{\mathbb{I}(Z_i=X_i)} (1-q)^{\mathbb{I}(Z_i \neq X_i)}}{q^{\mathbb{I}(Z_i=X'_i)} (1-q)^{\mathbb{I}(Z_i \neq X'_i)}} \\
&= \max \left(\frac{q^{\mathbb{I}(0=X_i)} (1-q)^{\mathbb{I}(0 \neq X_i)}}{q^{\mathbb{I}(0=X'_i)} (1-q)^{\mathbb{I}(0 \neq X'_i)}}, \frac{q^{\mathbb{I}(1=X_i)} (1-q)^{\mathbb{I}(1 \neq X_i)}}{q^{\mathbb{I}(1=X'_i)} (1-q)^{\mathbb{I}(1 \neq X'_i)}} \right)
\end{aligned}$$

taking also the maximum between $(X_i = 1, X'_i = 0)$ and $(X_i = 0, X'_i = 1)$, we obtain the constraint

$$\max \left(\frac{q}{1-q}, \frac{1-q}{q} \right) \leq \exp(\alpha)$$

therefore,

$$\frac{1}{1 + \exp(\alpha)} \leq q \leq \frac{\exp(\alpha)}{1 + \exp(\alpha)}.$$

The mutual information $I(X_{1:n}, Z_{1:n})$ is

$$\begin{aligned}
I(X_{1:n}, Z_{1:n}) &= \sum_{X_{1:n} \in \{0,1\}^n} \sum_{Z_{1:n} \in \{0,1\}^n} P(X_{1:n}, Z_{1:n}; p, q) \log \left(\frac{P(X_{1:n}, Z_{1:n}; p, q)}{P(X_{1:n}; p) P(Z_{1:n}; q, p)} \right) \\
&= \sum_{i=1}^n \sum_{x \in \{0,1\}} P(X_i = x) \left(\sum_{z \in \{0,1\}} P(Z_i = z | x) \log \frac{P(Z_i = z | x)}{P(Z_i = z)} \right) \\
&= n((1-p)q \log \frac{q}{(1-q)p + q(1-p)} + (1-p)(1-q) \log \frac{1-q}{pq + (1-q)(1-p)} \\
&\quad + p(1-q) \log \frac{1-q}{(1-q)p + q(1-p)} + pq \log \frac{q}{pq + (1-q)(1-p)}) \\
&= n f_p(q)
\end{aligned}$$

Taking the first two derivatives of f with respect to q

$$f'_p(q) = -\log\left(\frac{1}{q} - 1\right) + (1-2p) \log\left(\frac{1}{p+q(1-2p)} - 1\right)$$

and

$$f''_p(q) = \frac{(1-p)p}{q(1-q)[p+q(1-2p)][1-p-q(1-2p)]} \geq 0.$$

Since $f'_p(1/2) = 0$, we can deduce that f_p is decreasing on $[0, 1/2]$, reaches the minimum at $q = 1/2$, for which $X_{1:n}$ and $Z_{1:n}$ are independent, then is increasing. Therefore, including the differential privacy constraint, and noticing that $f(q) = f(1-q)$, we find that the two optimal q achieving the minimal amount of privacy required are

A formal approach to data swapping and disclosure limitation techniques

the ones on the boundary, $q^* = \frac{e^\alpha}{1+e^\alpha}$ or $q^* = \frac{1}{1+e^\alpha}$. Therefore the optimal swapping matrix is

$$M^* = \begin{bmatrix} q^* & 1-q^* \\ 1-q^* & q^* \end{bmatrix}$$

3.2 Extension to categorical key variable

Let us now consider the case of swapping a key variable with n_k possible outcomes, e.g. the geographical variable. $X_i \in \{1, \dots, n_k\}$ is now a Multinomial with probabilities (p_1, \dots, p_{n_k}) . Therefore the Marginal distributions of the sample $X_{1:n}$ is

$$P(X_{1:n}; p_{1:n_k}) = \prod_{i=1}^n \prod_{j=1}^{n_k} p_j^{\mathbb{I}(X_i=j)}. \quad (5)$$

We consider the swapping matrix

$$M = \begin{bmatrix} q_1 & \frac{1-q_1}{n_k-1} & \frac{1-q_1}{n_k-1} & \dots & \frac{1-q_1}{n_k-1} \\ \frac{1-q_2}{n_k-1} & q_2 & \frac{1-q_2}{n_k-1} & \dots & \frac{1-q_2}{n_k-1} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{1-q_K}{n_k-1} & \frac{1-q_K}{n_k-1} & \frac{1-q_K}{n_k-1} & \dots & q_{n_k} \end{bmatrix}$$

therefore, we obtain the conditional distribution of $Z_{1:n}$ given $X_{1:n}$ as follows

$$Q(Z_{1:n} | X_{1:n}, q_{1:n_k}) = \prod_{i=1}^n q_{X_i}^{\mathbb{I}(Z_i=X_i)} \prod_{j \neq X_i} \left(\frac{1-q_{X_i}}{K-1} \right)^{\mathbb{I}(Z_i=j)}. \quad (6)$$

The marginal of $Z_{1:n}$ is

$$P(Z_{1:n}; q_{1:n_k}, p_{1:n_k}) = \prod_{i=1}^n \prod_{j=1}^{n_k} \left(p_j q_j + \sum_{k \neq j} p_k \frac{1-q_k}{n_k-1} \right)^{\mathbb{I}(Z_i=j)},$$

while the joint distribution $P(X_{1:n}, Z_{1:n}; p, q) = Q(Z_{1:n} | X_{1:n}; q) P(X_{1:n}; p)$ is simply obtained by multiplying (6) and (5).

Let $X_{1:n}$ and $X'_{1:n}$ be s.t. $H(X_{1:n}, X'_{1:n}) = 1$, i.e. $X_l \neq X'_l$ and $X_i = X'_i$ for all $i \neq l$. Suppose $X_l = k$ and $X'_l = k'$ with k and k' different. Therefore condition (3) becomes

$$\begin{aligned}
\sup_{S \in \sigma(\mathcal{Z}^n)} \frac{\mathcal{Q}(Z_{1:n} \in S | X_{1:n})}{\mathcal{Q}(Z_{1:n} \in S | X'_{1:n})} &= \sup_{S \subseteq [1:K]^n} \sum_{Z_{1:n} \in S} \frac{\prod_{i=1}^n q_{X_i}^{\mathbb{I}(Z_i=X_i)} \prod_{j \neq X_i} \left(\frac{1-q_{X_i}}{n_K-1} \right)^{\mathbb{I}(Z_i=j)}}{\prod_{i=1}^n q_{X'_i}^{\mathbb{I}(Z_i=X'_i)} \prod_{j \neq X'_i} \left(\frac{1-q_{X'_i}}{n_K-1} \right)^{\mathbb{I}(Z_i=j)}} \\
&= \sup_{S \subseteq [1:n_K]^n} \sum_{Z_{1:n} \in S} \left(\frac{(n_K-1)q_k}{1-q_{k'}} \right)^{\mathbb{I}(Z_1=k)} \left(\frac{1-q_k}{(n_K-1)q_{k'}} \right)^{\mathbb{I}(Z_1=k')} \prod_{j \notin \{k,k'\}} \left(\frac{1-q_k}{1-q_{k'}} \right)^{\mathbb{I}(Z_1=j)} \\
&= \max \left(\frac{(n_K-1)q_k}{1-q_{k'}}, \frac{1-q_k}{(n_K-1)q_{k'}}, \frac{1-q_k}{1-q_{k'}} \mathbb{I}(K \geq 3) \right).
\end{aligned}$$

Hence we need to take the maximum over all possible pairs (k, k') in $[1 : n_K]$ with $k \neq k'$ to obtain the set of constraints in the optimization problem (4). Finally it can be seen that the mutual information $I(X_{1:n}, Z_{1:n})$ equals

$$\sum_{i=1}^n \sum_{x=1}^{n_K} p_x \left(\sum_{z=1}^{n_K} q_x^{\mathbb{I}(z=x)} \prod_{j \neq x} \left(\frac{1-q_x}{n_K-1} \right)^{\mathbb{I}(z=j)} \log \left(\frac{q_x^{\mathbb{I}(z=x)} \prod_{j \neq x} \left(\frac{1-q_x}{n_K-1} \right)^{\mathbb{I}(z=j)}}{p_z q_z + \sum_{k \neq z} p_k \frac{1-q_k}{n_K-1}} \right) \right).$$

In the setting outlined here, the solution of the optimization problem (4) is not available in closed form and computational methods are required: this will be the subject of future developments.

References

1. Agresti, A.: Categorical Data Analysis. 3rd Edition. Wiley, (2013).
2. Bethlehem, J.G., Keller, W.J., Pannekoek, J.: Disclosure control of microdata. J. Amer. Statist. Assoc. **85**, 38–45 (1990)
3. Dwork., C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In Proc. of the Third Theory of Cryptography Conference. pg 265–284 (2006)
4. Dwork., C., Roth, A.: The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science . **9**(3-4), 211–407 (2014)
5. Gibbs, A.L., Su, F.E.: On Choosing and Bounding Probability Metrics. Int. Stat. Rev. **70**, 419–435 (2002)
6. Gray, R.M.: Entropy and Information Theory. 2nd Edition. Springer, (2011).
7. Matthews, G.J., Harel, O.: Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. Statist. Surv. **5**, 1-29 (2011)
8. Rinott, Y., O’Keefe, C.M., Shlomo, N., Skinner, C.: Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. Statist. Sci. **33**, 358–385 (2018)
9. Shlomo, N., Skinner, C.J.: Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata. Ann. of App. Stat. **4**(3), 1291–1310 (2010)
10. Skinner, C.J., Elliot, M.J.: A Measure of Disclosure Risk for Microdata. J. Roy. Statist. Soc. B **64**, 855–867 (2002)
11. Skinner, C., Marsh, C., Openshaw, S., Wymer, C.: Disclosure control for census microdata. J. Off. Stat. **10**, 31–51 (1994)

A new ordinary kriging predictor for histogram data in L^2 -Wasserstein space

Un nuovo predittore kriging per istogrammi nello spazio L^2 -Wasserstein

Antonio Balzanella and Antonio Irpino and Rosanna Verde

Sommario This paper introduces an ordinary kriging predictor for histogram data. We assume that the input data is a set of histograms which summarize data observed in a geographic area. Our aim is to predict the histogram of data in a spatial location where it is not possible to get records. We consider the histograms as random elements of a L^2 -Wasserstein space. The isometry from the Wasserstein space of probability measures to the subset of $L^2(0, 1)$ of quantile functions, allows us to introduce a linear predictor which uses the quantile functions associated with the histograms.

Sommario *Questo articolo introduce un predittore kriging per dati ad istogramma. Si assume che i dati di partenza siano degli istogrammi che sintetizzano dati osservati in un'area geografica. Il nostro obiettivo è quello di prevedere l'istogramma dei dati in una posizione spaziale in cui non è possibile acquisire osservazioni. Gli istogrammi sono considerati come elementi casuali di uno spazio L^2 -Wasserstein. L'isometria dallo spazio di Wasserstein delle misure di probabilità al sottoinsieme di $L^2(0, 1)$ delle funzioni quantile, ci permette di introdurre un predittore lineare che usa le funzioni quantili associate agli istogrammi.*

Key words: Histogram data, kriging prediction, variogram for histogram data. . .

Antonio Balzanella
Università della Campani Luigi Vanvitelli, e-mail: antonio.balzanella@unicampania.it

Antonio Irpino
Università della Campani Luigi Vanvitelli, e-mail: antonio.irpino@unicampania.it

Rosanna Verde
Università della Campani Luigi Vanvitelli, e-mail: rosanna.verde@unicampania.it

1 Introduction

Nowadays, many real-world applications produce huge amounts of data which demand for high storage and computational resources. A common practice to address this issue is to summarize data before making the analysis through aggregates, i.e. averages of clusters of data, or by functional data, estimates of probability distributions, time series representation tools.

In this paper we consider geographic data summarized by histograms, that is our input data is a set of histograms each one having a spatial location. We can consider as examples of applicative scenarios the case of monitoring pollutions over time, at different locations of a geographic area; monitoring crimes in different cities; monitoring economical variables over time in different areas. In these examples we can use histograms as estimators of probability distributions which keep a detailed view of the data, reducing memory occupation and supporting fast computation. Since histograms record information about the moments of the data as well as the quantiles they are a more informative tool than simpler aggregates.

We assume that the covariance among histograms depend on their spatial location, consistently with Tobler Law[5] which states that "everything is related to everything else, but near things are more related than distant things". That is, we expect that histograms at close locations are more similar than histograms at far locations.

The aim of this paper is to introduce a kriging predictor which allows to predict the histogram of an unobserved spatial location as a weighted average of histograms, where weights depend on the spatial dependence among data.

We consider histograms as (estimates of) probability measures in the L^2 -Wasserstein space. The isometry from the Wasserstein space of probability measures to the subset of $L^2(0, 1)$ of quantile functions, allows us to introduce a linear predictor which uses the quantile functions of histograms.

2 Notations and main definitions

Let $H = \{H_1, \dots, H_i, \dots, H_n\}$ be a set of histograms defined as:

$$H_i = \{(I_{i,1}, \pi_{i,1}), \dots, (I_{i,l}, \pi_{i,l}), \dots, (I_{i,L}, \pi_{i,L})\}$$

where: $I_{i,l}$ are the L consecutive intervals (bins), with the associated weights $\pi_{i,l} \geq 0$, summing to 1.

We consider the set H as random elements of the L^2 -Wasserstein space.

We recall that Wasserstein spaces are metric spaces connected to the problem of optimal mass transportation (see [6] for a review). In this sense, the distance between elements of such space measures the minimal effort required to reconfigure the probability mass of one distribution in order to recover the other distribution.

Given two probability measures μ and ν on \mathfrak{R}^d with finite p -th moment, the Wasserstein distance of order p is defined as:

$$d_W(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathfrak{R}^d \times \mathfrak{R}^d} \|x - y\|^p d\pi(x, y) \right)^{1/p} \quad (1)$$

where $\Pi(\mu, \nu)$ denotes the set of transport plans between μ and ν , i.e. the set of probability measures on $\mathfrak{R}^d \times \mathfrak{R}^d$ having μ and ν as marginals.

The Wasserstein distances $d_W(\mu, \nu)$ are proper distances, since they are non-negative, symmetric, and satisfy the triangle inequality.

Closed expressions for Wasserstein distances d_W are available only in few cases such as that of probabilities on the real line: given two probability measures μ and ν on \mathfrak{R} with right-continuous distribution functions F and G , the Wasserstein distance of order p is defined as

$$d_W(\mu, \nu) = \left(\int_0^1 \|F^{-1}(t) - G^{-1}(t)\|^p dt \right)^{1/p} \quad (2)$$

where $F^{-1}(t)$ and $G^{-1}(t)$ ($t \in [0, 1]$) are left-continuous nondecreasing functions on $(0, 1)$ named quantile functions, obtained as the inverse of F and G . In other words, as shown in [4], $\mu \mapsto F^{-1}$ and $\nu \mapsto G^{-1}$ is an isometry from the Wasserstein space of the measures to the subset of $L^2(0, 1)$ formed by (equivalence classes of) left-continuous nondecreasing functions on $(0, 1)$.

A very important case is the Wasserstein distance of order $p = 2$, also known as quadratic Wasserstein distance. Such distance plays a central role among Wasserstein distances, just as L^2 plays a central role in the family of L^p spaces. A main feature of the $p = 2$ case is the possibility to consider a notion of barycentre of probability measures through the Fréchet mean.

By analogy with the Euclidean case where the barycenter of a set $x_1, \dots, x_i, \dots, x_n$ of points is obtained as minimizer of $x \mapsto \sum_{i=1}^n \lambda_i \|x_i - x\|^2$ (with weights λ_i), in [1] the same procedure is proposed in the Wasserstein space by simply replacing the squared Euclidean distance, with the squared L^2 -Wasserstein distance:

$$\mu_* = \inf_{\mu} \sum_{i=1}^n \lambda_i d_W^2(u_i, \mu) \quad \lambda_i \geq 0, \sum_{i=1}^n \lambda_i = 1 \quad (3)$$

where u_i and μ are probability measures with second finite moment.

No explicit solution to the Fréchet mean problem is available, in general, in the multivariate case, however, when $d = 1$, due to the flat nature of the Wasserstein space in such a case, the problem of Fréchet mean admits a simple and explicit solution since the L^2 -Wasserstein metric induces the $L^2(0, 1)$ geometry on the quantile functions of the measures [7]. As shown in [4], the Fréchet mean is the unique measure μ^* whose quantile function is the average of the quantile functions $\{F_i^{-1}, \dots, F_n^{-1}\}$ associated with the measures μ_i, \dots, μ_n .

By considering histograms as non-parametric estimates of probability density functions on the real line, the squared L^2 -Wasserstein distance between the histograms H_i, H_j is defined as:

$$d_W^2(H_i, H_j) = \int_0^1 (Q_i(\xi) - Q_j(\xi))^2 d\xi \quad (4)$$

where $Q_i(\xi)$ and $Q_j(\xi)$ are the quantile functions associated with the histograms H_i, H_j , respectively.

3 Ordinary kriging for histogram data

In the previous section, we have seen that the L^2 -Wasserstein metric induces a $L^2(0, 1)$ geometry on the quantile functions of probability measures on the real line. We consider quantile functions associated with histograms H_i (with $i = 1, \dots, n$) as random elements of a spatial random process χ_s .

Formally, let $\{\chi_s : s \in D \subseteq \mathbb{R}^d\}$ be a spatial random process, where s is a spatial location in a d -dimensional Euclidean space and D is a subset of \mathbb{R}^d with positive volume. We choose n points $s_1, \dots, s_i, \dots, s_n$ in D to observe the random functions $\chi_{s_i}(\xi)$, with $i = 1, \dots, n$. We assume that each $\chi_{s_i}(\xi)$, with $\xi \in (0, 1)$, is a random element of a subset of L^2 Hilbert space formed by left-continuous non-decreasing functions on $(0, 1)$, that is, quantile functions, equipped with a Borel σ -algebra and satisfying $\int_0^1 \chi_{s_i}(\xi)^2 d\xi < \infty$.

Further, χ_s is a second order stationary and isotropic spatial process, that is:

- $E(\chi_s(\xi)) = m(\xi)$, for $s \in D$
- $Cov(\chi_{s_i}(\xi_1), \chi_{s_j}(\xi_2)) = C(h, \xi_1, \xi_2)$, for $s_i, s_j \in D$, where $h = \|s_i - s_j\|$.

We define the variogram $\eta(h)$ for histogram data as follows:

$$\eta(h) = \frac{1}{2} E \left[\int_0^1 (\chi_{s_i}(\xi) - \chi_{s_j}(\xi))^2 d\xi \right] \text{ for } s_i, s_j \in D, h = \|s_i - s_j\|, \quad (5)$$

where $\int_0^1 (\chi_{s_i}(\xi) - \chi_{s_j}(\xi))^2 d\xi$ is the squared L^2 -Wasserstein distance between histograms H_i, H_j computed through the corresponding quantile functions χ_{s_i}, χ_{s_j} .

Consistently with [3], $\eta(h)$ is estimated by:

$$\hat{\eta}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \int_0^1 (\chi_{s_i}(\xi) - \chi_{s_j}(\xi))^2 d\xi. \quad (6)$$

As the classic variogram, $\eta(h)$ is nonnegative, $\eta(0) = 0$, and it is conditionally negative-definite.

Now, we introduce a linear predictor which uses the quantile functions Q_i of the histograms H_i . The predicted quantile function \hat{Q} is piecewise linear and it is uniquely associated with the histogram \hat{H} .

We consider the following linear predictor:

$$\hat{Q}(\xi) = \sum_{i=1}^n \lambda_i Q_i(\xi), \quad (7)$$

where:

$\hat{Q}(\xi)$ is the piecewise quantile function to predict;

$Q_i(\xi)$ is the piecewise quantile function associated with H_i ;

$\lambda_i \in \mathfrak{R}$ are the kriging weights.

Consistently with the concept of Fréchet mean in the L^2 -Wasserstein space of measures on \mathfrak{R} , the histogram predicted at the location s_0 is a linear combination of the quantile functions of the histograms H_i (with $i=1, \dots, n$) at the locations s_1, \dots, s_n . Due to the spatial dependence, the kriging weights λ_i in Eq. 7 are such that the locations which are closer to the prediction point have a greater influence than farther ones.

To ensure that $\hat{Q}(\xi)$ is a Best Linear Unbiased Predictor (BLUP) for $Q(\xi)$, the weights should satisfy the following conditions:

$$E \left[\int_0^1 (\hat{Q}(\xi) - Q(\xi)) d\xi \right] = 0$$

;

$$E \left[\int_0^1 (\hat{Q}(\xi) - Q(\xi))^2 d\xi \right] \text{ is minimum}$$

.

However, to constrain \hat{Q} to be a quantile function, the n weights must be non negative, thus the following constrained optimization problem has to be solved:

$$\min E \left[\int_0^1 \left(\sum_{i=1}^n \lambda_i Q_i(\xi) - Q(\xi) \right)^2 d\xi \right] \text{ s.t. } \sum_{i=1}^n \lambda_i = 1, \lambda_i \geq 0 \quad (8)$$

where $\sum_{i=1}^n \lambda_i = 1$ is the unbiasedness constraint and $\lambda_i \geq 0$ is the convexity constraint which ensures that \hat{Q} is still a quantile function.

The problem in Eq. 8 cannot be solved using classical Lagrange multipliers because of the inequality constraints. In [2], authors propose a solution to the problem of kriging with non-negative weights based on the Kuhn-Tucker theorem.

Following this approach, we assume that the $n \times n$ matrix η expressing the spatial dependence among sites, is defined as:

$$\boldsymbol{\eta} = \begin{pmatrix} \eta(\|s_1 - s_1\|) & \cdots & \eta(\|s_1 - s_i\|) & \cdots & \eta(\|s_1 - s_n\|) \\ \vdots & & \ddots & & \vdots \\ \eta(\|s_i - s_1\|) & \cdots & \eta(\|s_i - s_i\|) & \cdots & \eta(\|s_i - s_n\|) \\ \vdots & & \vdots & & \vdots \\ \eta(\|s_n - s_1\|) & \cdots & \eta(\|s_n - s_i\|) & \cdots & \eta(\|s_n - s_n\|) \end{pmatrix}$$

Thus, the weights $\lambda_1, \dots, \lambda_n$, minimizing the non-negative kriging objective function, are the solution to the following system of linear equations:

$$\begin{bmatrix} \boldsymbol{\eta} & \mathbf{1} & -\mathbf{I} \\ \mathbf{1}' & \mathbf{0} & \mathbf{0}' \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{u} \\ \boldsymbol{\delta} \end{bmatrix} = \begin{bmatrix} \eta_{s_0} \\ \mathbf{1} \end{bmatrix}$$

where:

- \mathbf{I} is an $n \times n$ identity matrix;
- $\boldsymbol{\lambda}$ is the n dimensional vector of kriging weights;
- $\mathbf{0}$ is an n dimensional vector of zeros;
- $\mathbf{1}$ is an n dimensional vector of ones;
- $\boldsymbol{\delta}$ is the n dimensional vector of the Lagrange multipliers associated with the non-negativity constraint;
- u is the Lagrange multiplier associated with the unbiased condition.

In [2], authors provide an efficient algorithm for finding an optimal solution.

The prediction variance as a global uncertainty measure, is given by:

$$\sigma_{s_0}^2 = \int_0^1 V \left(\widehat{\mathcal{Q}}(\xi) - \mathcal{Q}(\xi) \right) d\xi = \sum_{i=1}^n \lambda_i \eta(\|s_0 - s_i\|) - \mu. \quad (9)$$

Under a specified variogram model, we can use $\sigma_{s_0}^2$ to identify those zones for which we have greater uncertainty on the predictions.

4 Conclusions

In this paper we have introduced a kriging predictor for histogram data. Similarly to classic kriging, it allows to predict data at a geographic location where there is no possibility to get a measure of the analyzed phenomenon. However, differently from classic kriging, the prediction is not a scalar but a histogram. To address this challenge we have considered histograms as elements of a L^2 -Wasserstein space and we have developed the predictor consistently with the geometry of such a space. Future work will be the development of a detailed testing procedure to assess the effectiveness of the method.

Riferimenti bibliografici

1. Agueh M, Carlier G (2011) Barycenters in the Wasserstein space. Soc. Ind. Appl. Math. 43:904-924.
2. Barnes RJ, Johnson TB (1984) Positive kriging. In Geostatistics for natural resources characterization, pp. 231–244. Springer.
3. Delicado P, Giraldo R, Comas C, Mateu J (2010) Statistics for spatial functional data: some recent contributions. Environmetrics, 21(3–4), 224–239.
4. Panaretos VM, Zemel Y (2016) Amplitude and phase variation of point processes. The Annals of Statistics, 44(2):771–812.
5. Tobler W (1970) A computer movie simulating urban growth in the Detroit region. Economic Geography, 46(2), 234–240.
6. Villani C (2003) Topics in optimal transportation. Graduate Studies in Mathematics, 58. American Mathematical Society, Providence, RI.
7. Zemel Y, Panaretos VM. (to appear) Fréchet Means and Procrustes Analysis in Wasserstein Space. Bernoulli Journal.

Keywords dynamics in online social networks: a case-study from Twitter

*La dinamica delle parole chiave nelle reti sociali online:
un esempio tratto da Twitter*

Carolina Becatti, Irene Crimaldi and Fabio Saracco

Abstract This work presents a model that describes the users' activity in an online social network context. More precisely, it describes the evolution over time of the bipartite network formed by tweets and the hashtags they show. New tweets can include new hashtags or establish links with the already existing ones. The probability to form links with the old tags is regulated by a rule called "preferential attachment with weights". We apply this model to a dataset of tweets related to the period of the Italian elections of last March 2018 and discuss the obtained results.

Abstract *Il modello presentato descrive l'attività degli utenti in un contesto di reti sociali online. Più nello specifico, descrive l'evoluzione nel tempo della rete bipartita formata dai tweets e dagli hashtags in essi contenuti. I nuovi tweets possono introdurre nuovi hashtags nel sistema oppure contenere alcuni dei tags già presenti: la probabilità con cui questo accade è espressa in funzione di un meccanismo chiamato "preferential attachment con pesi". Il modello viene applicato ad un dataset di tweets relativi al periodo delle elezioni italiane dello scorso marzo 2018; presentiamo, quindi, i risultati di tale analisi.*

Key words: keywords dynamics, online social networks, Twitter, generative model, preferential attachment with weights, users' activity.

1 Introduction

During the last years we have witnessed an incredible revolution in terms of data availability: large amount of unstructured data are at our hand in short time and coming from totally different contexts. To this extent, the use of networks [6] has revealed extremely powerful in shaping and extracting information from these

Carolina Becatti, Irene Crimaldi, Fabio Saracco
IMT School for Advanced Studies Lucca, Piazza San Ponziano 6, 55100 Lucca, Italy
e-mails: carolina.becatti@imtlucca.it, irene.criminaldi@imtlucca.it, fabio.saracco@imtlucca.it

sources. Two branches of network theory, that are also the main categories this work refers to, deal with generative models [3] and lexical networks [7, 8]. The former describe the temporal evolution of an observed graph: starting from an initial set of connected nodes, new nodes are added at every following time-step; the newcomers form links with the already existing nodes according to a pre-defined rule. The latter instead consist of a set of words connected to each other by a certain relation, such as being synonyms or co-occurrence or phonological similarity.

This work models the activity in the online social network of Twitter, describing the evolution over time of the bipartite network (i.e. a network which set of nodes can be partitioned in two subsets and only links across different subsets can be observed) formed by tweets and the hashtags they show. It is based on the model presented in [1], that in turn finds its roots in [4, 5]. New posts can bring new hashtags, not present in the system by the time of their arrival, or establish links with the already existing ones. The probability to form links with the old tags is regulated by a rule called “preferential attachment with weights”: it is a function of the hashtag popularity but also depends on some individual features (*fitnesses*, in general) of the user posting the tweet, the tag involved, or the posted tweet.

2 The model

Consider a system of $i = 1, \dots, n$ users that sequentially post some content on the social network. Notice that, as in the model presented in [1], the sequence of time stamps coincides with the flow of tweets, therefore the term “time-step t ” indicates the time in which tweet t has been posted.

This model describes the dynamical evolution of the bipartite network that collects users’ tweets and the corresponding hashtags. We indicate with F the adjacency matrix related to this graph. The dynamics starts with the observation of the first tweet that shows N_1 hashtags, where N_1 is assumed Poisson distributed with parameter $\alpha > 0$ ($\text{Poi}(\alpha)$). We enumerate the observed tags from 1 to N_1 and we set $F_{1,h} = 1$ for $h = 1, \dots, N_1$. Then, for each following post $t \geq 2$, we have:

1. Tweet t shows some of the hashtags already present in the system, exhibited by one of the published posts $\{1, \dots, t-1\}$. More precisely, if N_j denotes the number of new tags shown by tweet j and $L_{t-1} = \sum_{j=1}^{t-1} N_j$ indicates the number of different hashtags observed with the first $t-1$ tweets, the new post t can independently display each old tag $h \in \{1, \dots, L_{t-1}\}$ with probability

$$P_t(h) = \frac{\delta}{2} + (1 - \delta) \frac{\sum_{j=1}^{t-1} F_{j,h} W_{t,h}}{t} \quad (1)$$

where $\delta \in [0, 1]$ is one of the model parameters, $F_{j,h} = 1$ if tweet j shows tag h and $F_{j,h} = 0$ otherwise, $W_{t,h} \geq 0$ is a random weight, associated to hashtag h , measured at the time of post t . Finally we divide by t so that the ratio on the right

in equation (1) belongs to $[0, 1]$. We will refer to the quantity in equation (1) as the “inclusion probability” of hashtag h at time-step t .

2. Tweet t can also show a number of new hashtags N_t , where N_t is assumed $\text{Poi}(\lambda_t)$ -distributed with parameter

$$\lambda_t = \frac{\alpha}{t^{1-\beta}}$$

where $\beta \in [0, 1]$ is a parameter. The variable N_t is supposed independent of $\{N_1, \dots, N_{t-1}\}$ and of all the old hashtags and their weights, including those appeared at time t .

We indicate with \mathcal{H} the set of hashtags appeared at the end of the observation period T . There is a wide range of possible ways in which the weights can be defined. Refer to [1] for a more detailed discussion regarding the model’s parameters and the possible definitions of the weights.

3 Dataset

Using the Twitter API for Python, we have downloaded a sample of all tweets posted from January 28 to March 19, 2018. The API is a particularly useful tool, since it allows to automatically isolate special elements from the text (such as hashtags, urls, users’ mentions, media, etc.) without requiring the authors to perform expensive pretreatment analyses on the data. The selection process is performed by means of a list of keywords: each post must contain at least one of a set of elections-related keywords in the Italian language, such as *elezioni*, *elezioni2018*, *4marzo*, *4marzo2018*. For the following analysis we have considered only the tweets receiving at least one retweet and containing at least three hashtags. Moreover, we have also excluded the retweets from the construction of the features matrix, since they could represent a source of bias for the model, being an exact copy of the original post. Finally, because of the large amount of activity in online social networks during the election days, in few cases we observe different tweets that have been posted at the same time. To overcome this limitation, when constructing the features matrix we order those tweets alphabetically according to the tweet ID number in string format. The final number of tweets in our sample is reported in the first column of Table 1.

4 Choices of the weights

The analysis starts with the selection of three definitions of weights to model the inclusion probability in equation (1). In this work the weights are expressed as:

- Type 1: a decreasing function of the number of time-steps intervening between the current time t and time t_h^* , that is the last time hashtag h appeared. Older tags, that showed up some time before t , may be less popular and therefore less likely

Table 1 Number of tweets and hashtags available after the data preparation procedure.

	overall	C ₁	C ₂	C ₃
tweets	5742	2696	1939	2371
hashtags	2827	1563	1132	1378

to be used in new tweets.

$$W_{t,h} = e^{-G_{t,h}} \quad \text{with} \quad G_{t,h} = t - t_h^* - 1 \quad \text{and} \\ t_h^* = \max\{j : 1 \leq j \leq t - 1 \text{ and } F_{j,h} = 1\} \quad (2)$$

- Type 2: an increasing function of the total number of tags tweeted by the user who posts the tweet appeared at time t . The higher/lower the number of tags posted by an user in her/his previous posts, the more/less incline she/he is in adopting hashtags in the future.

$$W_{t,h} = W_t = e^{-1/G_{i(t)}} \quad \text{with} \quad i(t) = \text{author of tweet } t \\ G_{i(t)} = \text{number of hashtags tweeted by user } i(t) \text{ until time } t - 1 \quad (3)$$

- Type 3: a decreasing function of the number of followers of an user. The higher the number of followers, the higher the popularity of an account may be. Therefore it may be reasonable to think that she/he is less affected by other people's posts and less incline to use "old" hashtags.

$$W_{t,h} = W_t = e^{-G_{i(t)}} \quad \text{with} \quad i(t) = \text{author of tweet } t \\ G_{i(t)} = \text{number of followers of user } i(t) \text{ at time } T \quad (4)$$

5 Results

As a first step we split the initial sample of all tweets in three different subsets $\mathcal{C} = \{C_1, C_2, C_3\}$ in order to divide the activity by discussion topics on the Italian political coalitions or parties. The selection process has been performed using the words in Table 4: each tweet is claimed to be discussing about the community $C \in \mathcal{C}$ if it contains at least one of the hashtags listed in the corresponding column. The final number of tweets and hashtags related to each subset is reported in Table 1. Notice that with this mechanism the division in \mathcal{C} does not form a partition of the overall set of tweets and a subset of 375 tags are common to the three groups. In addition to that, since we do not analyse the tweet text, both tweets supporting and criticizing a certain coalition may belong to it: if the hashtag #berlusconi appears in a tweet, then the discussion is focused on him, independently if in support or against him. Then the model has been run separately on the three subsets.

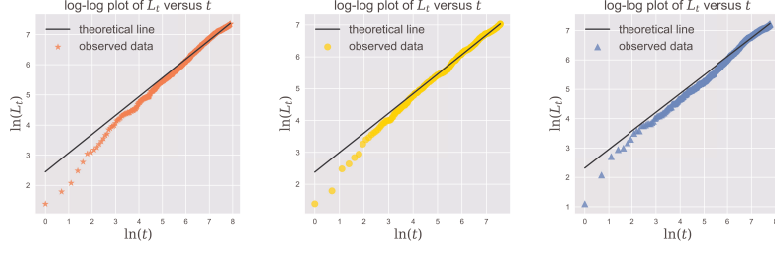


Fig. 1 Plot of $\ln(L_t)$ as a function of $\ln(t)$ with power-law trend. The coloured points refer to the real data while the black line provides the theoretical regression line with slope $\hat{\beta}$.

Table 2 shows that the parameters are unbiasedly estimated from the data with the three considered kinds of weights. Figure 1 provides a graphical representation of the cumulative count of hashtags as a function of time in the three subsets: the observed quantity shows the same power-law behaviour proven for the model, as shown by the black line with the theoretical slope $\hat{\beta}$ (see [1]). Moreover the preferential attachment plays a relevant role in the dynamics, because of the very small values observed for the parameter δ . In this regard, there is a difference between the three weights: the value of δ estimated for the case with weights in (4) is systematically higher than in the other cases and this difference is preserved across the three subsets of tweets.

Table 3 shows instead a comparison between the observed networks and the simulated ones, with the three considered kinds of weights and with flat weights (i.e. all weights equal to 1). The total number of hashtags L_T and the average number of new hashtags \bar{N}_T are well reproduced by the four alternatives. The major differences are in the average number of old hashtags \bar{O}_T : this quantity is clearly underestimated by the model with the weights in (2) and overestimated by the model with flat weights. There is instead an agreement between the observed quantity and the estimated one with the weights (3) and (4): in both cases, the tweets shows an average number of three hashtags previously appeared in the dynamic and this result is in line with the observed network.

6 Concluding remarks

This work describes the temporal evolution of the activity in an online social network context: we model the process of link formation in the bipartite network of tweets and the hashtags they show. The results of our analyses indicate that the power law behaviour of the total number of hashtags in time perfectly matches the theoretical one, with exponent equal to the model parameter β . Moreover the small value obtained for the parameter δ indicates that the preferential attachment rule is an important driver for this dynamics. However, the presence of the weights ex-

Table 2 Estimation of the model parameters for each considered weight-types. See the Appendix, equation (5) for a description of these quantities.

C_1	\hat{p}	\bar{p}	$MSE(p)$	C_2	\hat{p}	\bar{p}	$MSE(p)$
α	7.1434	7.1300	0.6858	α	6.6885	6.7069	0.7012
β	0.6227	0.6237	0.0003	β	0.6143	0.6145	0.0004
δ weights in (2)	0.0034	0.0034	$2.67 \cdot 10^{-9}$	δ weights in (2)	0.0046	0.0046	$6.39 \cdot 10^{-9}$
δ weights in (3)	0.0039	0.0039	$4.37 \cdot 10^{-9}$	δ weights in (3)	0.0060	0.0060	$1.23 \cdot 10^{-8}$
δ weights in (4)	0.0071	0.0071	$5.02 \cdot 10^{-9}$	δ weights in (4)	0.0103	0.0103	$1.52 \cdot 10^{-8}$

C_3	\hat{p}	\bar{p}	$MSE(p)$
α	6.3970	6.3973	0.6019
β	0.6331	0.6337	0.0003
δ weights in (2)	0.0042	0.0042	$3.79 \cdot 10^{-9}$
δ weights in (3)	0.0045	0.0045	$5.98 \cdot 10^{-9}$
δ weights in (4)	0.0082	0.0082	$7.98 \cdot 10^{-9}$

Table 3 Comparison between real and simulated matrices by means of the indicators in equation (6) in the Appendix for each considered weight-types.

C_1	L_T	σ_{L_T}	\bar{O}_T	σ_{O_T}	\bar{N}_T	σ_{N_T}
real	1563		3.561		0.580	
weights in (2)	1559.90	38.69	1.655	0.082	0.579	0.014
weights in (3)	1562.36	38.70	3.234	0.115	0.580	0.014
weights in (4)	1565.98	39.60	3.434	0.102	0.581	0.015
weights = 1	1563.51	41.54	16.181	2.228	0.580	0.015

C_2	L_T	σ_{L_T}	\bar{O}_T	σ_{O_T}	\bar{N}_T	σ_{N_T}
real	1132		3.641		0.584	
weights in (2)	1128.06	32.69	1.665	0.091	0.582	0.017
weights in (3)	1130.34	34.38	3.344	0.133	0.583	0.018
weights in (4)	1133.74	34.19	3.593	0.128	0.585	0.018
weights = 1	1131.09	35.19	14.627	2.244	0.584	0.018

C_3	L_T	σ_{L_T}	\bar{O}_T	σ_{O_T}	\bar{N}_T	σ_{N_T}
real	1378		3.571		0.581	
weights in (2)	1379.48	36.75	1.802	0.107	0.582	0.016
weights in (3)	1376.40	36.73	3.251	0.118	0.581	0.016
weights in (4)	1377.27	37.04	3.447	0.113	0.581	0.016
weights = 1	1377.84	36.06	14.510	2.048	0.581	0.015

pressed as a function of a fitness variable is fundamental for reproducing the real data. Indeed, in most of the considered cases, the model that takes the weights into account performs better in matching the the number of “new” and “old” hashtags and the total number of tags (refer to equation (6) for their definition).

Despite this fact, there is a difference in the performance of the three definitions of weights: the worse behaviour of the weights in (2) denotes that this quantity is not a good proxy for the importance of a word. This is reasonable, since the online dynamic is very quick, many tweets are posted every minutes on Twitter; since the timesteps of our model coincides with the published tweets, many time instants intercurring from the last appearance of a word may, instead, indicate only minutes,

in terms of the real-world time. Therefore this may not be a good indication of how popular an old word is and how likely it is used. On the contrary, the weights in (3) and (4) guarantee a much better agreement with the observed data. Therefore both these fitnesses may be considered good representatives for, respectively, the tendency of an author to use tags and the level of popularity and influence of an user in the considered context. On the one hand, hashtags are tools used in online social networks to attract attention and gain popularity, therefore only those users who post on Twitter with this purpose may be incline to choose the most appropriate old tags and include them in their tweets. On the other, the accounts that have a large number of followers are often famous people, newspapers and parties accounts, political figures, that especially in the elections period are only interested in widening their pool of voters and spreading their ideas [2]. Therefore they might be interested in using the tags that best reproduce them, regardless of whether these words have been used before. One possible future development of this work includes the application of the model to the tweets of the users classified in political alliances as in [2]: the accounts have not been manually classified but their division in clusters is the result of the users' retweeting behaviour.

Appendix

Table 4 List of hashtags used to select the tweets belonging to each discussion topic.

Group	Tags
Left leaning (C_1)	pd, partitodemocratico, renzi, leu, liberieuguali, piueuropa, poterealpopolo, pc, partitocomunista, civicapolare, matteoreenzi, grasso, pietrograsso, bonino, emmabonino, lorenzin
Movimento 5 Stelle (C_2)	m5s, votom5s, votom5stelle, movimento5stelle, beppegrillo, luigidimaio, dimaio, toninelli, danilotoninelli
Right leaning (C_3)	forzaitalia, fratteditalia, lega, casapound, salvinipremier, noiconlitalia, forzanuova, italiaagliitaliani, salvini, matteosalvini, meloni, giorgiameloni, berlusconi, silvioberlusconi

The model parameters are estimated using the tools provided in the Appendix of [1]. Then, we simulate the model dynamics generating $R = 500$ realizations of the tweets-hashtags network. For each parameter $p \in \{\alpha, \beta, \delta\}$, we compute the average value \bar{p} and the mean squared error $MSE(p)$ over the set of R realizations, as follows

$$\bar{p} = \frac{1}{R} \sum_{r=1}^R \hat{p}_r \quad \text{and} \quad MSE(p) = \frac{1}{R} \sum_{r=1}^R (\hat{p}_r - \bar{p})^2 \quad (5)$$

where \hat{p} is the parameter estimate while \hat{p}_r is the value obtained for the r -th realization of the model. In order to compare the simulated dynamics with the observed one, we compute the following indicators

$$L_T = \text{total number of hashtags appeared in the observed } T \text{ tweets}$$

$$\bar{O}_T = \frac{1}{T-1} \sum_{t=2}^T O_t \text{ with } O_t = \sum_{h=1}^{L_{t-1}} F_{t,h} \text{ and } \bar{N}_T = \frac{1}{T} \sum_{t=1}^T N_t. \quad (6)$$

The quantities \bar{O}_T and \bar{N}_T provide respectively the average number of “old” and “new” hashtags overall the set of tweets. We compute these quantities on the observed matrix and on $R = 500$ simulations of the dynamics, with and without the selected weights. In particular, for each indicator $I \in \{L_T, \bar{O}_T, \bar{N}_T\}$ the tables report the average value \bar{I} and the sample standard deviations σ_I , as follows

$$\bar{I} = \frac{1}{R} \sum_{r=1}^R I_r \text{ and } \sigma_I = \sqrt{\frac{1}{R-1} \sum_{r=1}^R (I_r - \bar{I})^2} \quad (7)$$

where I_r is the indicator I computed over the r —th simulation of the model.

Acknowledgements

CB and IC are members of the Italian Group “Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro Applicazioni” (GNAMPA) of the Italian Institute “Istituto Nazionale di Alta Matematica” (INdAM).

Funding

FS is supported by the EU projects CoeGSS (Grant No. 676547), Openmaker (Grant No. 687941), SoBigData (Grant No. 654024). IC and FS are supported by the Italian “Programma di Attività Integrata” (PAI), project “TOol for Fighting FakeEs” (TOFFE) funded by IMT Lucca.

References

1. Becatti, C., Crimaldi, I., Saracco, F.: Collaboration and followership: a stochastic model for activities in social networks. Preprint arXiv:1811.00418 (2018)
2. Becatti, C., Caldarelli, G., Lambiotte, R., Saracco, F.: Extracting significant signal of news consumption from social networks: the case of Twitter in Italian political elections. Preprint arXiv:1901.07933 (2019)
3. Bianconi, G., Barabási, A. L.: Competition and multiscaling in evolving networks. Europhys. Lett. **54**(4), 436 (2001)
4. Boldi, P., Crimaldi, I., Monti, C.: A network model characterized by a latent attribute structure with competition. Inf. Sci. **354**, 236–256 (2016)
5. Crimaldi, I., Del Vicario, M., Morrison, G., Quattrociocchi, W., Riccaboni, M.: Modelling networks with a growing feature-structure. Interdiscip. Inf. Sci. **23**(2), 127–144 (2017)
6. Newman, M. E. J.: Networks. An introduction. Oxford University Press, Oxford (2010)
7. Stella, M.: Modelling early word acquisition through multiplex lexical networks and machine learning. Big Data Cogn. Comput. **3**(1), 10 (2019)
8. Stella, M., Beckage, N.M. and Brede, M.: Multiplex lexical networks reveal patterns in early word acquisition in children. Sci. Rep. **7**, 46730 (2017)

Statistical Matching of HBS and ADL to analyse living conditions, poverty and happiness

Statistical Matching di HBS e ADL per l'analisi di condizioni di vita, povertà e felicità

Cristina Bernini, Silvia Emili, Maria Rosaria Ferrante

Abstract Consumption, poverty and happiness represent fundamental aspects in the analysis of household living conditions. To empirically investigate the relationships among them, individual data are required. The not availability of joint information on consumption and happiness at the unit level, as in Italy, may be overcoming by using the statistical matching method. In particular, the matching of the Household Budget Survey (HBS) with the Aspects of Daily Life (ADL) provides information at the individual level, useful to investigate how poverty as well as the living condition affects the happiness of Italian citizens.

Abstract *Consumi, povertà e felicità costituiscono aspetti fondamentali nell'analisi delle condizioni di vita di una comunità. Al fine di investigare le relazioni che intercorrono tra esse, sono necessari dati puntuali sui diversi aspetti ma osservati congiuntamente con un dettaglio individuale. La non disponibilità di dati ad un livello fine di disaggregazione, come è il caso in Italia, può essere superata grazie all'utilizzo di metodi di statistical matching. Il matching dell'Indagine sui Bilanci delle famiglie (HBS) e della Multiscopo (ADL) consente di investigare come la povertà e le condizioni di vita sono influenzate dalla felicità degli italiani.*

Key words: statistical matching, synthetic data source, living conditions, poverty, happiness

¹ Cristina Bernini, Department of Statistical Sciences "P. Fortunati", University of Bologna; Center for Advanced Studies on Tourism (CAST), University of Bologna, Rimini; email: cristina.bernini@unibo.it

Silvia Emili, Department of Statistical Sciences "P. Fortunati", University of Bologna; Center for Advanced Studies on Tourism (CAST), University of Bologna, Rimini; email: silvia.emili2@unibo.it

Maria Rosaria Ferrante, Department of Economics, University of Bologna; email: maria.ferrante@unibo.it

1 Introduction

One of the goals of the literature on subjective well-being¹ is the relationship between happiness and income. Several studies have underlined that richer individuals (or countries) are happier than poorer individuals (or countries) (Clark, 2017); that is, with respect to income, richer individuals are more satisfied with their lives (Diener et al., 2010). The negative relationship between poverty and happiness is more significant in the upper segments of the income distribution. In addition, there is evidence that self-reported life satisfaction is lower for those who are classified as being in poverty (Clark et al., 2015). Besides, people may be unhappy about poverty even if they are not poor themselves; they may feel bad about the possibility of becoming poor themselves in the future. They may also feel repercussions on the economy and society as a whole (Welsh et al, 2019).

We extend the poverty-happiness literature by focusing on different measure of poverty and happiness. Specifically, we consider not only satisfaction with the overall life but also the satisfaction expressed by individuals in respect to the different life domains (i.e. relationships with relatives, friends, environment, economic condition, leisure time). As for poverty, we suggest considering not only to be in poverty but also the gap to the poverty line. However, this information is not jointly available from the same source; thus, we suggest reconstructing a novel database containing information on consumption as well as happiness at the individual level by means of a statistical matching tool. The statistical matching procedure employed in this paper answers to the need of obtaining information about individual consumptions and different aspects of happiness and quality of daily life. The joint availability of micro-data on these dimensions gives alternative advantages on the evaluation of poverty and consumer behaviours on different expenditure groups, on the analysis of territorial disparities in well-being and on the investigation of the multidimensionality of happiness for alternative expenditure categories. In particular, the last aspect motivates our interest in obtaining a synthetic dataset starting from two ISTAT's survey: the Household Budget Survey (HBS) and the Aspects of Daily Life (ADL) survey. In the analysis, we suggest matching the two surveys by using a non-parametric micro approach, as the distance hot deck technique (Okner, 1972).

Using a data set of 35,975 individuals in 2016, obtained by matching the two surveys, we show that Italian people happiness (i.e., measured by overall satisfaction or life domains satisfaction) is negatively correlated with poverty measures (i.e. poverty status and poverty intensity).

2 The Data

¹ As Veenhoven (2012) argues, subjective well-being “it is an umbrella term for all that is good. In this meaning, it is often used interchangeably with terms like ‘well-being’ or ‘quality of life’ and denotes both individual and social welfare”. Accordingly, we will use the terms ‘subjective well-being’, ‘happiness’ and ‘life satisfaction’ interchangeably.

Statistical matching of HBS and ADL to analyse living conditions, poverty and happiness

The empirical analysis is performed using two data sources: for individual data, we exploit the “Multipurpose Survey on Households: Aspects of Daily Life” (ADL). Expenditure data are extracted from the Household Budget Survey (HBS). All information is provided by the Italian Office of Statistics (ISTAT).

ADL is a large repeated cross-sectional sample survey that covers the resident population in private households, collecting annual information on individual and household daily life. A sample of 43,251 residents (18,508 households) in Italy was collected in 2016 and observed with respect to both life satisfaction and residents’ socio-demographic characteristics. Life satisfaction is measured through the question: “At this moment, how much are you satisfied with your life overall?” Respondents may answer on an 11-point Likert- scale, where 0 is the lowest and 10 the maximum level of satisfaction. The Likert scale is standard in measuring interpersonal satisfaction (Diener and Seligman, 2002, *inter alia*). Overall, the mean life satisfaction equals 7.03 in 2016 with a right-skewed distribution, which reveals quite a high level of happiness among citizens. Together with an overall measure of happiness, the ADL observes citizens’ satisfaction with respect to the different domains of their life, by using a 4-point Likert scale. This set includes: the relationship with relatives (average value: 3.24) as well as friends (3.05) and health condition (2.95), activities during leisure time (2.77), satisfaction for the economic and environmental conditions (2.42 and 2.81 respectively). The satisfaction aspects are integrated with a large set of individual characteristics, related to the socio-demographic characteristics of residents (gender, age, marital status, household composition, educational level), economic condition (in search of occupation, retirement) and both household and individuals’ habits.

The HBS survey covers the resident population in private households (35,975 individuals in 15,409 households), collecting information about the expenditure habits. In particular, information on monthly consumption behaviour of households for a wide set of durable commodities (e.g. dwelling products), as well as non-durables (e.g. food, clothing, health), and services, differentiating with respect to periodic and occasional consumption. General information on family’s components (demographic characteristics and information about dwelling) completes the survey. A relative poverty indicator could be calculated as the base of the household total expenditure, excluding unusual expenses, allowing the analysis of distribution and intensity of poverty. Differently from ADL, HBS represents one of the most used Italian survey in statistical matching framework (Conti et al., 2017).

3 The statistical matching

Statistical matching, also known as data fusion or data integration (for a review see among others: D’Orazio et al., 2006b), represents one of most popular statistical tools for combining data taken from different official statistics. The specific literature field reveals a long story characterized by peaks and troughs of interest (Kadane, 1978; Rubin, 1986; Moriarity and Scheuren, 2001, 2003; Rassler, 2002; D’Orazio et al., 2005, 2006a; Conti, et al., 2016; 2017). One of the main

classifications of these techniques refers to the final purpose of the procedure, and it distinguishes between micro approaches (developed in order to provides a final dataset as synthetic file containing all disjoint information about the underlying population), and macro approaches (focused on direct estimation of joint distribution function or of key measures of the disjoint observed target variables).

Formally, statistical matching frameworks can be summarize referring to the presence of two independent sample survey, A and B, which records are considered as independently and identical generated from appropriate models. The two samples can be represented by

$$A = A_n = \{(Y_i, X_i)\}_{i=1}^n, \quad B = B_m = \{(X_j, Z_j)\}_{j=1}^m,$$

implying Z missing in A's units and Y missing in B, X present in both samples (i.e., matching variables), generated from $f(x, y, z)$, with $F = \{f\}$ be a suitable family of density. Even if we can avoid the (implicit) assumption of a specific estimate of the distribution (D'Orazio et al., 2006b), a natural way to avoid misleading assumptions on F is achieved using nonparametric matching procedures. By this way, the class of hot deck methods seem strongly interesting and appealing.

The most appropriate setting of matching variables X_M has been defined as compromise between the subset of variables that better explain Y and Z, i.e. X_Y and X_Z , with $X_Y \subseteq X$, and $X_Z \subseteq X$, and is given by

$$X_Z \cap X_Y \subseteq X_M \subseteq X_Z \cup X_Y.$$

A nontrivial issue in statistical matching applications, relates to the conditional independence assumption (CIA). This problem concerns independence of the target variables Y and Z conditioning on the common variable X. Even if CIA cannot be tested, in the micro approach this problem is overcome by introducing auxiliary information (correlation coefficients as well as additional datasets).

Starting from early papers of Okner (1972), distance hot deck procedure appears as one of the most used micro approach in empirical applications, due in part to the its intuitive way of dealing with statistical matching problems: the imputation of a live value from the donor file (that provides values for missing observation in the second sample) to the recipient (that will receive imputed live values for the missing observations) according to some distance (similarity measures) between common variables (henceforth *matching variables*). Moreover, the constrained version of the distance hot deck procedure, guarantees the possibility to relate units between the two datasets only once (i.e. without replacement). This approach considers the identification of donor classes in which the number of donors must be equal or greater than the number of recipients, ensuring the preservation of the marginal distribution of the imputed variables. Beyond its straightforward applicability a second motivation underlying the use of distance hot deck techniques relates to Marella et al. (2008). The authors show that the matching noise due to distance hot deck decreases as the sample size of the donor sample increases.

As far our analysis, we start by identifying of the donor and the recipient datasets: with more than 40 thousand statistical units in ADL and about 36 thousand in HBS, the first survey is identified as the donor file. To complete the specification of the constrained distance hot deck problem, we need to define donor classes and

Statistical matching of HBS and ADL to analyse living conditions, poverty and happiness matching variables. Aiming at preserving marginal distribution of imputed data, we consider a three-level macro area variable as donation class, while we consider X_M given by: *Age13* (13-levels variable about age), *Ncomp* (number of household components), *Edu* (categorical variables stating for education), *mStat* (marital status), *Rgn* (20 Italian regions), and two variables referred to household disposable income, *Disp* (4 levels), and household economic status with respect to previous year, *Inc* (5 levels). The aim of the last two matching variables pertains to the use of information that could relax CIA. In this case, the information collected through *Disp* and *Inc*, can be considered as data concerning the relationship between Y and Z: the variables not only pertains to the definition of an overall satisfaction level, but (i) refer to changes in a general household income level, (ii) embodies temporal information about perceived income changes as temporary or permanent and hence about consumption attitude of the family.

The coherence of the synthetic data set is investigated comparing marginal distributions of imputed variables, where reference distribution is observed in the donor dataset. In particular, to evaluate the representativeness, we refer to four empirical measures: Bhattacharyya coefficient, Hellinger distance, total variation distance and the overlap between the two distributions (D’Orazio et al., 2006b).

Table 1: Comparison of marginal distributions

	<i>Battacharyya</i>	<i>Hellinger</i>	<i>Tvd</i>	<i>Overlap</i>
<i>LS</i>	0.999999	0.000609	0.000525	0.999475
<i>LDeco</i>	0.999999	0.000808	0.000874	0.999126
<i>LDenv</i>	0.999999	0.000915	0.001090	0.998910
<i>LDrel</i>	0.999987	0.003583	0.003937	0.996063

Table 1 shows marginal distribution comparisons about target variable (overall happiness, hereafter LS) and a subset of life domain satisfaction scores: satisfaction about economic condition (LDeco), environment (LDenv) and family relationships (LDrel). Results display a good preservation of the target variable in the matched file, with total variation (tvd) and Hellinger distance values closed to 0, and remaining indicators (i.e. Battacharyya and Overlap) closed to 1.

4 Modelling poverty and happiness of Italian citizens: some preliminary results

The aim of the analysis is to investigate to what extent being poor affects the individual satisfaction with respect to the overall life and to the different life domains. We estimate micro-econometric life satisfaction regressions in which overall happiness or life domain satisfaction of individual i depends on poverty measures and a set of individual-level controls. As poverty measures, we follow the official ISTAT practice and classify individual as poor (*dPoor*) if his/her equivalent consumption is below the poverty line (z), defined through Carbonaro's equivalence

scale (Istat, 2018). The scale is applied to transform each household disposable consumption into the household disposable equivalent consumption. Moreover, to investigate the intensity of the phenomenon on poor and non-poor individual, we define the poverty ratio (PG) for the i -th unit ($i=1, \dots, n$) as the individual equivalent consumption x_i with respect to the poverty line:

$$PG_i(x, z) = \frac{x_i}{z}$$

By this way, a value lower/greater than 100 stands for individual with expenditures below/beyond the poverty line.

From this starting point, the final model specification is given by:

$$\ln(LS) = f(dPoor, PG, W)$$

where $\ln(\cdot)$ is the natural logarithm, $f(\cdot)$ is a suitable continuously differentiable linear function, $dPoor$ is a dummy variable classifying individuals as poor, PG is the poverty ratio, $W = (W_{ind}, W_{hou})$ contains individuals and household variables, in particular $dFemale$ is a dummy variables for female, $dMacroArea_k$, $k=1,2,3,4,5$ is a set of dummies for Italian macro area (North west, North east, Center, South, Island), $dComponent_j$, $j=1,2,3,4$, is a dummy variable for the number of household components (from 1 to 4, with 4 stands for households with 4 or more individuals), $AgeMid$ is the midpoint of age classes in Age13, and $dEducation_h$, $h=1,2,3,4,5$, is a set of dummy variables related to highest education level (from 1 to 5, no education, primary school, secondary school, high school, university), $Rooms$ represents the number of dwelling rooms. Model estimates are carried out over the whole matched-sample and on the subset of poor people (i.e, people lying under the poverty line); in both samples, we model either the overall or the life domain satisfaction.

As preliminary results, we show OLS estimates in Table 2 and Table 3 about overall satisfaction and life domain scores, respectively. In line with the literature, we find positive and statistically significant association between overall satisfaction score and expenditure measures in the overall sample, and negative sign with the poverty dummy. Nonetheless, highest magnitudes can be associated to education level and number of household components. Excluding non-poor individuals, the coefficient associated to PG strongly increases, highlighting the enhancement about the importance of increases in expenditure levels on happiness.

Considering life domain satisfaction, and after having controlled for W , the impact of poverty on happiness related to the economic situation, environment condition and relationships with relatives (Table 3) confirm previous results, with minor exceptions. Being poor has a negative and relevant effect on the satisfaction with own economic condition; while the impact on other domains are lower and negligible. The relative position with respect to the poverty line mainly affects the economic condition especially in the sub-sample of the poor. The lowest effect of being poor and distant from the poverty line has been detected for the satisfaction with relatives.

Table 2: Overall satisfaction estimates (***) $p < 0.01$, ** $p < 0.05$, * $p < 0.1$)

Statistical matching of HBS and ADL to analyse living conditions, poverty and happiness

	<i>Overall Sample</i>		<i>Poors Sample</i>	
<i>Const</i>	1.724***	(0.04)	1.541 ***	(0.12)
<i>dPoor</i>	-0.017**	(0.01)	-	-
<i>Ln(PG)</i>	0.024***	(0.01)	0.077 ***	(0.03)
<i>dFemale</i>	0.012***	(0.00)		
<i>dMacroArea₁</i>	0.030***	(0.01)		
<i>dMacroArea₂</i>	0.027***	(0.01)	-0.072 ***	(0.03)
<i>dMacroArea₄</i>	-0.017***	(0.01)	-0.068 *	(0.02)
<i>dMacroArea₅</i>	0.003	(0.01)	-0.043 **	(0.02)
<i>dComponent₁</i>	-0.036***	(0.01)	-0.079 ***	(0.02)
<i>dComponent₃</i>	-0.019***	(0.01)		
<i>dComponent₄</i>	-0.014**	(0.01)		
<i>Ln(AgeMid)</i>	-0.017***	(0.01)		
<i>dEducation₂</i>	0.062***	(0.01)		
<i>dEducation₃</i>	0.078***	(0.01)		
<i>dEducation₄</i>	0.106***	(0.01)	0.068 ***	(0.02)
<i>dEducation₅</i>	0.126***	(0.01)		
<i>Ln(Rooms)</i>	0.022***	(0.01)		

Table 3: Life domain satisfaction estimates of poverty variables (*** p<0.01, ** p<0.05, * p<0.1)

<i>Independent</i> <i>Dependent</i>	<i>Overall Sample</i>		<i>Poors Sample</i>	
	<i>dPoor</i>	<i>Ln(PG)</i>	<i>Ln(PG)</i>	
<i>LS</i>	-0.017*** (0.01)	0.024*** (0.01)	0.077*** (0.03)	
<i>LDeco</i>	-0.039*** (0.00)	0.077*** (0.01)	0.187*** (0.03)	
<i>LDenv</i>	-0.005 (0.01)	0.023*** (0.00)	0.053 ** (0.03)	
<i>LDrel</i>	-0.011* (0.01)	0.009** (0.00)	0.017 (0.02)	

5 Preliminary conclusions and further developments

In the analysis of living conditions, poverty and happiness of Italian citizen, the unavailability of a complete and exhaustive survey about main features of the phenomenon is encompassed resorting to statistical matching of HBS and ADL. The analysis of data from the synthetic file shown preliminary results, in line with economics of happiness literature. However, due to endogeneity and matching noise, models relating satisfaction to expenditure or poverty can be affected by inconsistency when estimated through ordinary least squares. As for future research, we will consider instrumental variables estimator (IV) to investigate the relationship between happiness and poverty from the synthetic dataset. Further robustness

procedures about preliminary results are then implemented referring to two-sample two stage least squares and bias correct estimators of Hirukawa and Prokhorov (2018). Moreover, other future development of this research regards the modelling of the effect of being in poverty on a multidimensional composite indicator of life satisfaction, constructed on the base of the different life domains.

References

1. Clark, A. E. Happiness, Income and Poverty. *Int Rev Econ.* 64, 145 – 158 (2017)
2. Clark, A. E., Flèche, S., Senik, C. Economic growth evens out happiness: Evidence from six surveys. *Rev. Income Wealth* (2015)
3. Conti, P.L., Marella, D., Neri, A. Statistical Matching and Uncertainty Analysis in Combining Household Income and Expenditure Data. *Stat Methods Appl.* 26:3, 485-505 (2017)
4. Conti, P.L., Marella, D. Scanu, M. Statistical Matching Analysis for Complex Survey Data with Applications. *J Am Stat Assoc.* 111:516, 1715-1725 (2016)
5. DeLeire, T., Kalil, A. Does consumption buy happiness? Evidence from United States. *Int Rev Econ.* 57, 163 – 176 (2010)
6. Diener, E.; Seligman, M. Very Happy People. *Psychol Sci.* 13, 81 – 84 (2002)
7. Diener, E., Ng, W., Harter, J., Arora, R. Wealth and Happiness Across the World: Material Prosperity Predicts Life Evaluation, Whereas Psychological Prosperity Predicts Positive Feeling. *J Pers Soc Psychol.* 99, 52– 61 (2010)
8. D’Orazio, M., Di Zio, M. e Scanu, M. Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. *J Off Stat*, vol. 22, n. 1, pp. 1-12 (2006a)
9. D’Orazio M., Di Zio M., and Scanu M.: *Statistical Matching, Theory and Practice.* Wiley, New York (2006b)
10. Hirukawa, M. Prokhorov A. Consistent Estimation of Linear Regression Models Using Matched Data. *J Econometrics.* 203(2), 344 - 358 (2018)
11. Istat. La povertà in italia - Anno 2017, Statistiche Report (2018)
12. Kadane, J.B. Some Statistical Problems in Merging Data Files, Reprinted in 2001 in *J Off Stat*, 17, 423-433 (1978)
13. Marella, D., Scanu, M., Conti, P. L. On the Matching Noise of Some Nonparametric Imputation Procedures. *Stat Probabil Lett*, 78, 1593-1600 (2008)
14. Moriarity C., Scheuren F. Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure. *J Off Stat.* 17, 407-422 (2001)
15. Moriarity C., Scheuren F. A Note on Rubin's Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputation, *J Bus Econ Stat.* 21, 65-73 (2003)
16. Okner, B.A. Constructing a new data base from existing microdata sets: the 1966 merge file. *Ann. Econ. Soc. Meas.* 1, 325-342 (1972)
17. Rassler S. *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches.* Springer Verlag, New York (2002)
18. Rubin D.B. Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *J Bus Econ Stat.* 4, 87-94 (1986)
19. Welsh, H., Biermann, P. Poverty is a Public Bad: Panel Evidence form Subjective Well-Being Data. *Rev Income Wealth.* 65(1), 187 - 200 (2019)

Statistical sources for cybersecurity and measurement issues

Fonti statistiche per la sicurezza cibernetica e problemi di misurazione

Claudia Biancotti, Riccardo Cristadoro, Raffaele Tartaglia Polcini

Abstract Growing awareness of cybersecurity as a critical requirement of today's economy raises the demand for a correct measurement of related risk. To date, however, reliable and well documented sources of information are scant while cyber incident statistics published by private firms, often with vested interests in the cybersecurity business, have fed the media with little control on their quality. Sound estimates are key to sound policy decisions and pricing of risk, which underpins insurance premia. A few bright spots are starting to appear, mainly within the advanced economies; global interconnectedness of cyber risk calls for international coordination and standardization.

Abstract *La crescente consapevolezza dell'importanza della sicurezza cibernetica per i sistemi economici solleva la questione della corretta misurazione del rischio. Finora, comunque, le fonti informative affidabili e ben documentate rimangono scarse; le statistiche sugli incidenti cyber pubblicate da imprese private, spesso in conflitto di interesse, affollano i media senza un reale controllo sulla loro qualità. Stime affidabili sono cruciali per decisioni politiche efficaci e per il risk pricing su cui vengono basati i premi assicurativi. Si intravedono miglioramenti, in particolare, nelle economie avanzate; l'interconnessione globale del rischio cyber richiede coordinamento e standardizzazione a livello internazionale.*

Key words: cybersecurity, statistical sources, measurement of risk, international coordination

¹ All authors: Bank of Italy, International Relations and Economics Directorate. Claudia Biancotti: claudia.biancotti@bancaditalia.it; Riccardo Cristadoro: riccardo.cristadoro@bancaditalia.it; Raffaele Tartaglia Polcini (corresponding author): raffaele.tartagliapolcini@bancaditalia.it. Usual disclaimers apply.

1 Introduction

Attention to the issue of *cybersecurity* started spreading beyond the narrow scope of IT specialists and managers worldwide as the fast growing digitalization of the whole economy (OECD 2014; Bank of Italy 2018) and the increased interconnectedness of data networks revealed the complex and comprehensive nature of cyber risk and the interdependency of seemingly unrelated critical infrastructures. The buzzword *cybersecurity*¹ may sound like a simple catchy rewording of the sterner “IT security”, but it implies a holistic consideration of all aspects of technological security, not limited to the individual operating unit, in a cross-sectoral, cross-country, cross-platform framework.

Cybersecurity emerged to the public attention as an impressive slew of attacks to public and private infrastructures showed the fragility of the existing controls, at the national and international level, while cyber incursions against a number of established multinational companies proved able to significantly dent their profitability and reputation, especially in case of data theft. Major incidents began to be counted and classified (see, for example, CSIS 2019); the issues of related costs and spillovers came to the foreground, as well as that of measurement of cyber risk as a specific type of risk.

Quantitative information on cyber risk and related measures is, to date, mainly provided by private entities, namely cybersecurity and consulting companies. These run their own surveys and publish data, seldom accompanied by methodological notes and very likely biased, possibly in order to feed hype on how urgent it is to tackle the cybersecurity issue. Although this urgency may be justified, the plethora of widely disseminated yet undocumented figures clearly exposes a need for data on which to base strategic decisions; but the provision of sound statistics on this new topic requires the involvement of official providers.

Persistent lack of reliable quantitative information is already feeding renewed fallacies in the evaluation of cyber risk, as assessed, for example, in Hubbard *et al.* (2016): referring to the findings of Kahneman and Tversky (1979), the authors describe the perils and limitations of experts’ judgement in the specific field of cyber risk.

Lack of official data on cybersecurity can be in itself considered as a vulnerability, in that it distorts the planning of defensive efforts. This lack has recently taken center stage in high level international fora, testifying both the concerns provoked by cyber threats and the key role played by reliable sources of information.

In May 2017 (under the Italian presidency), finance ministers and central bank governors of G7 countries acknowledged the existence of a data gap on the economic dimensions of cybersecurity; they called on “international organizations

¹ Widespread use of prefix *cyber-* to signify “IT-related”, with an emphasis on dangers and risks, seems to have originated from the warfare literature in the early 90s of 20th century.

Statistical sources for cybersecurity and measurement issues and governmental institutions in partnership with the private sector” to deliver “reliable, impartial, comprehensive and widely accessible” information (G7 2017) upon which policy decisions could be based.

2 National contributions

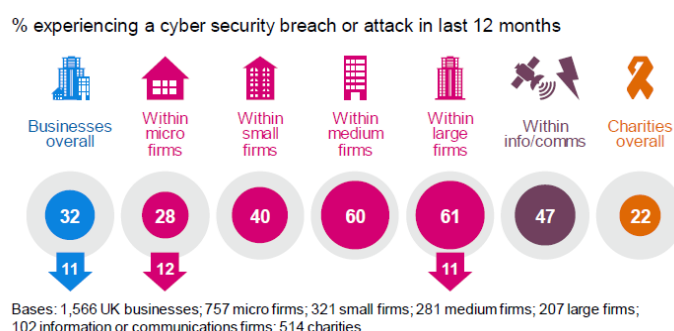
Pioneers in the field of official-source statistics data production for cybersecurity have been the United Kingdom and Italy.

2.1 *The UK Survey of Cyber Security Breaches*

One notable example of data of appropriate quality and transparency is the yearly UK Government’s Survey of Cyber Security Breaches, first conducted in 2016; the most recent results were published last April (UK Department for Culture, Media and Sport 2019). The availability of a few time points allows for an initial assessment of the cybersecurity developments over time: according to the survey, 32 per cent of businesses experienced cyber security breaches or attacks in the last 12 months. The overall proportion of breach occurrences is much lower than in 2018 (43 per cent) and 2017 (46 per cent). As in previous years, this percentage shows higher among medium and large businesses (60 per cent), but much lower than in previous year (71%). Although the drivers of these developments have yet to be investigated, the report points to a rise in cybersecurity level and awareness; on the other hand, mandatory compliance with the GDPR¹ from May 2018 might have changed what businesses define as a breach, or render some less willing to admit security breaches. The median number of reported breached has yet gone up, from 2 attacks in 2017 to 6 in 2019. The average cost of the breaches was £22,700 in 2019 for large firms. This is in line with corresponding value observed in 2018 (£22,300) and higher than 2017 (£19,600). Although these figures remain remarkably low, a rising trend seems to emerge. These results show clearly how useful the collection of detailed information can be for policy evaluation.

¹ The General Data Protection Regulation (EU) 2016/679 was implemented on 25 May 2018. According to the new rules, corporate personal data controller must put in place appropriate technical and organisational measures to implement the data protection principles set up in the law: lawfulness; fairness and transparency; purpose limitation; data minimisation; accuracy; storage limitation; integrity and confidentiality. The controller shall be responsible for, and be able to demonstrate compliance with, these principles (accountability). A specific provision of GDPR is that personal data breaches must be reported to the supervisory authority (Art. 33), “unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons”.

Fig. 1 Incidence of cyber attacks on UK firms, 2019



Source: UK Government.

2.2 The Bank of Italy cybersecurity survey of 2017

Italy has contributed to the big picture since the last two years. An exercise similar, if less detailed, to the UK was conducted by the Bank of Italy in 2017, leveraging the long-standing survey of Italian manufacturing and non-financial service firms. This initiative allowed, for the first time in Italy, to obtain reliable estimates on the behaviour of firms vis-à-vis cybersecurity and the impact of occurrences, both in terms of frequency and cost. A question on the attacks experienced has been administered also in 2018 and 2019.

In the questionnaire administered in 2017 (with reference to what happened in 2016) (Bank of Italy 2017) questions were aimed at measuring the existing controls and the related expenses, the frequency of cyber incidents, the damage undergone and their cost, the subsequent reaction on the management's side.

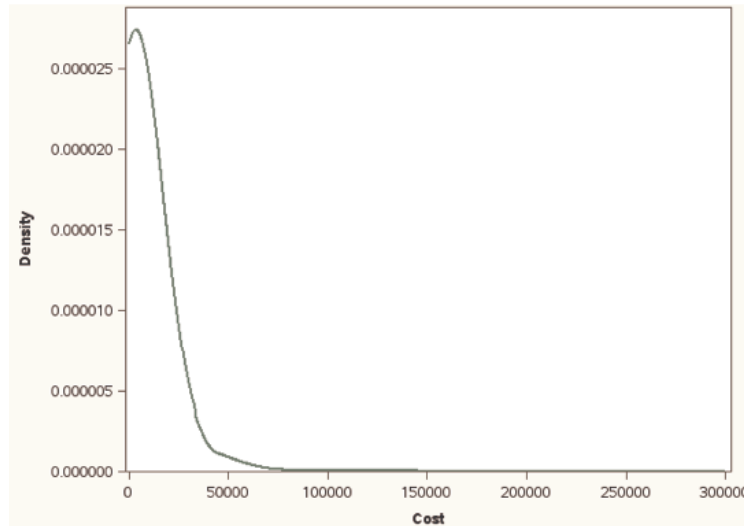
Reluctance to acknowledge cyber risk within corporate activities, as suggested in the UK survey, is analyzed and modeled for Italian data in Biancotti (2017a). Even though fewer than 2 per cent of businesses declared that they do not deploy any cybersecurity measures, about one third suffered at least some damage from a cyber attack between September 2015 and September 2016. An imputation model was developed to account for the unwillingness to report or the inability to detect attacks for some respondents; the share of firms attacked during the observed period climbed to 45.2, with large, high-tech and internationally exposed businesses faring worse than average.

Bank of Italy data also indicate that in 2016 the overwhelming majority of attacks against Italian nonfinancial private firms caused direct and recovery costs below €50,000, but one in a thousand victims reported damages of at least €200,000 (Biancotti 2017b). A growing body of evidence shows that the distribution of costs is highly asymmetrical, which limits the information content of most survey-based estimates. Edwards et al (2016) find that the size of breaches catalogued in the

Statistical sources for cybersecurity and measurement issues

Privacy Rights Clearinghouse dataset follows a log-normal distribution. This seems to be confirmed in the Bank of Italy survey (Fig. 2).

Figure 2. Monetary costs of all cyber attacks suffered in 2016, at the firm level (kernel density estimate; cost in euros)



Source: Bank of Italy, Survey of industrial and service firms, 2016. *Note:* respondents had the option of answering the cost question by choosing a bracket rather than indicating a point estimate. The upper bracket was open (“€ 200,000 and over”) and for purposes of readability it was capped at €300,000 in this chart. The figure should not be interpreted as an actual upper bound.

3 Some measurement issues

As seen in the Introduction, cyber risk measurement entails some specific features that are not entirely aligned with traditional operational risk and raise new measurement challenges. *Inter alia*:

- Evolving technology makes risk evaluation a moving target;
- The subject bearing the risk may not always coincide with the victims of the security failure; spillovers are often present;
- Risk should be measured not only at the individual level, but also at the group level (vaccine-like);
- Consequences of failure may show long after the breach, and/or last for a very long time;

- Some long lasting consequences of breaches and incidents are difficult to measure, typically damages to reputation;
- “Tail risk” (the combination of low probability and high impact) is particularly relevant¹.

These peculiarities suggest also that, in the case of cybersecurity, even traditional definitions of basic quantities like probability of occurrence and cost may prove inadequate. Measuring the simplest of variables, the frequency of incidents (to estimate the probability), is already quite challenging. Victims have scant incentives to disclose attacks, even in the presence of legal obligations; the reputational costs and subsequent loss of competitiveness can outweigh the benefits of information sharing (Gal-or and Ghose 2005, Laube and Böhme 2016).

When it comes to economic impact, the endeavour becomes significantly more difficult for conceptual and practical reasons. There is no shared definition of what constitutes the cost of a cyber attack. Most existing studies focus on the damage suffered by businesses that are directly hit by hackers, and even within this limited scope, some components are intrinsically uncertain. While the combined effect of the various distorting factors suggests that the cost of cyber attacks will never be estimated with the same precision achieved, say, for annual working hours, this is not necessarily a damning obstacle for policy design and evaluation. Selecting the right policy requires knowing which pieces of legislation, norms, and regulations were most successful in containing the economic impact of occurrences. If sources of measurement bias are well understood, it can be possible to derive unbiased estimators of variations in frequency and cost of attacks over time, even if the absolute levels cannot be pinned down exactly.

Improving the quality of cost indicators requires some groundwork in order to develop shared definitions of the economic cost of cyber attacks at various levels – individual, firm, sector, economy, group of economies – and common standards to measure those costs. Work is already underway in this area – the OECD established an international expert group, and the US Department of Homeland Security (Livingston et al. 2017; Department of Homeland Security 2017) recently proposed a measurement framework.

It is also important to pursue research on how best to integrate multiple data sources. Given the complexity of the phenomenon, it is unlikely that meaningful estimates can be obtained from a single dataset, no matter how good. Available surveys use the firm as a sampling unit. This serves the important purpose of showing the consequences of an attack on a firm’s economic performance, but it leaves the aggregation problem unsolved. Such surveys could be supplemented with others where incidents are the sampling unit; since no census of incidents exists, appropriate sampling frames have to be conjured from external sources – for example, incident notification archives managed by national data protection authorities, or lists of attacks such as the one maintained by the Center for Strategic and International Studies (2019).

¹ Some consider also the notion of *threat* as a low-probability, high-impact event, of which the probability is yet to be assessed.

Statistical sources for cybersecurity and measurement issues

Furthermore, given the importance of infrequent, large attacks, all data collection and estimation tools should be designed with rare events in mind. Oversampling the right tail, whenever one wants to measure a skewed, long-tailed variable, is common practice, where under-reporting and non-response are significantly higher compared to the rest of the distribution. Survey results can be usefully integrated with qualitative information on serious attacks – techniques adopted, speed of spreading across networks, time to detection, etc. Big data techniques can be leveraged to integrate heterogeneous, non-structured datasets covering the various dimensions of cybersecurity. The private sector is experimenting with these techniques (Lloyd’s 2017), and official statistics need to catch up.

4 International coordination and standardization: the financial system as a frontrunner

The strict interconnectedness of cyberspace - hardly bound by national borders, sectoral differences, platform diversity - and the freshness of most initiatives around the theme suggest that international coordination of activities, standardization of methods and definitions, integration of data sources could proceed fast, without frictions typical of internally homogeneous contexts. The financial system can be considered as a frontrunner, being a key target of cyber-attacks and having a high level of technological innovation and adoption. Attacks to financial infrastructures and intermediaries span a large set of motives, from profit to the intention to disrupt the orderly functioning of the economy, by means that can span from simple identity thefts to more sophisticated aggressions like APTs (advanced persistent threats)¹, the consequences of which can even reach a systemic level.

Central banks and supervisory authorities play a key role in ensuring the cybersecurity of the financial system. They manage and/or supervise vital components, such as payments systems, and are thus able to enforce collection of useful data on incidents by supervised external entities and adoption of suitable defence measures. With respect to other contexts, the long tradition of technical collaboration among financial authorities – especially within the Eurosystem – and their role in guaranteeing safety and security of transactions make this ecosystem very favourable for international cooperation, hence achieving better results.

A first step toward a shared view of the nature and extent of the threat and hence a common set of policies is a clear-cut set of terms and definitions, substituting the existing confusion of terms and meanings. To this end, in November 2018, the Financial Stability Board released a “Cyber Lexicon” defining a set of approximately 50 core terms, related to cyber security and cyber resilience in the financial sector. In December, the Basel Committee on Banking Supervision (BCBS) paved the way to international standardization and coordination of

¹ An intrusion that remains undetected for an extended period of time.

Claudia Biancotti, Riccardo Cristadoro, Raffaele Tartaglia Polcini
cybersecurity measures in the financial sector by publishing a report that describes and compares the range of existing bank, regulatory and supervisory cyber-resilience practices across jurisdictions. The report correctly points out that many “cyber-security and resilience metrics are not yet mature”, namely:

Some jurisdictions have methodologies to assess or benchmark regulated institutions’ cyber-security and resilience. Those jurisdictions that have developed ways to assess cyber-security and resilience have focused on reported incidents, surveys, penetration tests and on-site inspections. None of these methodologies produce quantitative metrics or risk indicators comparable to those available for financial risks and resilience, eg standardised quantitative metrics where established data are available. Instead, indicators provide information on regulated institutions’ approach to building and ensuring cyber-security and resilience more broadly. Supervisory authorities also rely on entities’ own management information, although this differs across entities and is not yet mature. (BCBS 2018, p.21)

Beyond constituting an essential instrument for cybersecurity policy evaluation and implementation, the provision of reliable data for dissemination should be part of a wider communication strategy aimed at raising awareness, also among the general public. In fact, guaranteeing the cybersecurity of firms, networks and infrastructure is not just a technological problem. Their vulnerability often stems from an undervaluation of the risks and from distorted incentives that lead to organizational failures, imprudent behaviour of staff, and insufficient investment in protection. Many attacks are carried out using simple tools that could easily be neutralized through compliance with basic rules of caution.

References

1. Bank for International Settlements and Board of the International Organization of Securities Commissions: 'Guidance on Cyber Resilience for Financial Market Infrastructures' (2016)
2. Bank of Italy: "Cyber-risk and the Italian economy". In: Bank of Italy Annual Report 2017, pp. 203-215. (2018)
3. –: Survey of Industrial and Service Firms in 2016 (2017).
<https://www.bancaditalia.it/pubblicazioni/indagine-imprese/2016-indagine-imprese/index.html>
4. Basel Committee on Banking Supervision: Cyber-resilience: Range of practices. <https://www.bis.org/bcbs/publ/d454.pdf> (2018)
5. Biancotti, C: 'Cyber Attacks: Preliminary Evidence from the Bank of Italy's Business Surveys', Occasional Papers no. 373, Bank of Italy (2017a)
6. –: "The price of cyber (in)security: Evidence from the Italian private sector", Bank of Italy, Occasional Papers no. 407 (2017b)
7. –, R Cristadoro, S Di Giuliomaria, A Fazio and G Partipilo: "Cyber attacks: An economic policy challenge", VoxEU.org, 23 June (2017)
8. –, R Cristadoro: "The machine stops: The price of cyber (in)security", VoxEU.org, 17 February (2018)
9. Center for Strategic and International Studies: 'Significant Cyber Incidents since 2006' (2019)
10. Department of Homeland Security (DHS): "National Cybersecurity Workforce Framework." <https://niccs.us-cert.gov/> (2017)
11. Edwards, B, S Hofmeyr and S Forrest: "Hype and heavy tails: A closer look at data breaches", Journal of Cybersecurity 2(1) (2016)
12. Financial Stability Board: Cyber Lexicon. <http://www.fsb.org/2018/11/cyber-lexicon/> (2018)
13. G7: G7 Finance Ministers and Central Banks' Governors Meeting Communiqué (2017)
14. Gal-or, E and A Ghose: "The economic incentives for sharing security information", Information Systems Research 16(2) (2016)
15. Hubbard, D W, R Seiersen, D E Geer: How to Measure Anything in Cybersecurity Risk. Wiley (2016)
16. Kahneman, D and A Tversky: Prospect Theory: An Analysis of Decision under Risk. Econometrica, Vol. 47, No. 2 (Mar., 1979)
17. Laube, S and R Böhme: "The economics of mandatory security breach reporting to authorities", Journal of Cyber Security 2(1): 29-41 (2016)
18. Livingston, O, M Shabbat and T Cheesebrough: "Cost of cyber incidents", presented at the 16th Workshop on the Economics of Information Security (2017)
19. Lloyd's: Counting the cost: Cyber exposure decoded (2017)
20. OECD: 'Measuring the Digital Economy. A New Perspective' (2017)
21. OECD: Enhancing the role of insurance in cyber risk management (2017)
22. UK Department for Culture, Media and Sport: 'Cyber Security Breaches Survey: Main Report' (2019)

Use of GPS-enabled devices data to analyse commuting flows between Tuscan municipalities

Un'analisi dei flussi di pendolarismo sistematici tra i comuni toscani tramite l'utilizzo di dati GPS

Chiara Bocci, Leonardo Piccini and Emilia Rocco

Abstract During the last two decades, we have seen an explosion in the deployment of pervasive systems like cellular networks, GPS devices, and WiFi hotspots that allow us to collect digital information about individual and collective behaviour of people at a very high level of geographical detail and can represent an extremely useful source of data. This is the case of a phenomenon like commuting whose identification and quantification at local level can produce a more detailed depiction of the socio-economic environment in which people live. In this paper we investigate to what extent big data collected from GPS-enabled devices, installed on private vehicles for insurance purposes could be a support in producing estimates of systematic commuting flows between municipalities in Tuscany.

Abstract Negli ultimi due decenni abbiamo assistito a un notevole incremento di strumenti digitali come reti cellulari, dispositivi GPS e hotspot WiFi in grado di produrre e memorizzare tracce digitali sul comportamento individuale e collettivo delle persone a un livello molto elevato di dettaglio territoriale e che rappresentano una fonte di dati estremamente utili per lo studio di numerosi fenomeni. Tra questi, il pendolarismo la cui identificazione e quantificazione a livello locale può produrre una rappresentazione dettagliata dell'ambiente socio-economico in cui le persone vivono. In questo articolo esaminiamo in che misura i big data raccolti da dispositivi GPS, installati su veicoli privati per scopi assicurativi, possono essere un utile fonte di dati per stimare i flussi di pendolari sistematici tra i comuni della Toscana.

Key words: Big data, commuting flows, Gravity models, Self-selection, Representativeness, Zero-Inflated Negative Binomial model

Chiara Bocci, Emilia Rocco

Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence
e-mail: chiara.bocci@unifi.it; emilia.rocco@unifi.it

Leonardo Piccini

IRPET - Regional Institute for Economic Planning of Tuscany e-mail: leonardo.piccini@irpet.it

1 Introduction

Technological evolution brought along, in recent years, a remarkable increase in the diffusion of devices that can record digital footprints of our behaviour on a daily basis, tracking a vast degree of activities. Constant and basically unintentional production of such tracks generates huge datasets that contain a precious quantum of information about socio-economic behaviour that may be extracted and used for research and for policy analysis.

Big data sources may support policy makers in the ex-ante phase of policy implementation, by providing a more sophisticated depiction of the socio-economic environment and may be used for ex-post evaluation purposes in quasi-experimental design and counterfactual settings. Literature on the matter and practical experiences have highlighted pros and cons of this approach. Some of the pros include timeliness, cost effectiveness, spatial and temporal disaggregation, emergence of unexpected and/or unobservable phenomena. On the other hand, since the relative novelty of the methodologies used to deal with these data, extra carefulness needs to be used to acknowledge possible shortcomings in terms of quality, accessibility, applicability, relevance, privacy policy and ownership of the data, all of which may affect the quality of policy evaluation and appraisal. Nonetheless, we believe that big data sources can be successfully used to foster the capabilities of the public institutions to deal with complex problems, to plan effective policies and to evaluate the outcomes of their actions.

Cities are fertile grounds for this kind of approach, as complex systems of people, things, environments, and activities. Digital signs that people produce every day by interacting with devices, social media, and other technological systems offer interesting opportunities to study and understand city dynamics and social behavior. The smart city paradigm provides a framework to try to understand these complex dynamics and deliver innovative solutions to meet citizens' needs (Bergamaschi *et al.*, 2016).

Big data has found in mobility analysis one of its most prolific fields of application. Traditionally, mobility analysis has suffered for lack of available and spatially detailed data. Data produced from widespread personal GPS and GSM devices can be therefore used for a more detailed analysis of mobility choices, in a more timely and efficient way, and may provide a very powerful tool for mobility planning, policy evaluation and economic appraisal of transport related investments (both in infrastructure and in services). To this extent, we propose a methodology that allows us to use data collected from GPS enabled devices, installed on private vehicles for insurance purposes, to analyse and understand mobility patterns.

2 Potential use of GPS-enabled devices: the case of study

The aim of the paper is to find a viable method to use GPS data to produce an estimate for the detailed Origin-Destination (OD) matrix between all pairs of municipi-

palities within Tuscany. Since the GPS dataset is derived from private car mobility, our primal focus is on this type of flows. However, in order to assess the intensity and characteristics of the relations between different geographic zones within the regional area, a future step of the analysis should be devoted to estimate public transport mobility patterns as well, possibly by considering the different propension on public transport usage which we expect to observe across the different OD pairs.

Typically, detailed OD matrix data are collected systematically every 10 years, during the nationwide official Census. However, census data, while very rich with information and details, has two major drawbacks: the temporal lag between censuses, during which we have no information on mobility, and the focus on what we call systematic mobility, *i.e.* the mobility which happens almost every day and is mainly related to home-to-school or home-to-work trips, leaving out an increasingly relevant segment of non-systematic mobility, which, by its own nature, is difficult to capture with traditional methods. Conversely, the Multipurpose Surveys on Household (MSH) carried out by Istat provides yearly estimates on commuting mobility patterns but only at regional level. In particular, it provides regional total estimates of the internal flows in all municipalities, of the outflows towards municipalities within the same province, of the outflows towards other provinces within the same region and of outflows to different regions, partitioned by means of commuting (car, train, bus, bike, etc...). Therefore, the MSH estimates are more frequent but less detailed than census data, and again focused only on systematic mobility.

If our methodology will be proved feasible, we can increase our analytical capability with an informative base that can be updated almost continuously and that includes all mobility and not only the systematic one.

For this study we use GPS data that are provided by a leader company in the Insurance Telematics that deals with about the 2% of the total vehicles circulating in Italy. Our dataset counts about 150k private vehicles crossing Tuscany in a period of six weeks between February and March 2014, and represents a primary source of information for studying the mobility behaviours. The history H of a user is represented by the set of points in space and time recorded by his mobility device; that is, $H = (p_1, \dots, p_n)$, where $p_i = (x, y, t)$, x, y are spatial coordinates and t is an absolute time-point. Starting from this sequence, we are interested in extracting the user's trajectories, where a trajectory is a sub-sequence of points representing the movement between two places in which the user stops for an activity. In the literature, there are several complex techniques for stop computation, yet in our experience a good trade-off between computational efficiency and quality of results can be obtained by a simple cut based on spatial and temporal thresholds; that is, a stop is detected (and therefore a trajectory ends and a new one starts) when the object remains within a region of radius r for at least a period of time γ . A common setup of the parameters in the case of private cars, tested in several analytical tasks, is $\gamma = 2\text{hours}$ and $r = 50\text{m}$ (Giannotti *et al.*, 2011). In the following, we will also use the concepts of "the most frequent location" L_1 and "the second most frequent location" L_2 . These two areas represent the two most visited places by the user, usually corresponding to *home* and *workplace* (Nanni, 2016). Therefore, for each user the systematic flows are identified as the flows observed between his L_1 and L_2 .

All the individual commuting flows are then aggregated at municipal level to obtain the OD matrix between the 269 Tuscan mainland municipalities (obtained considering the Tuscan administrative partition as existing in 2014 and removing the insular municipalities). Therefore, the municipal OD matrix comprises of 269 internal flows (the diagonal matrix elements) and 72092 outflows (the off-diagonal matrix elements), most of which will be zero-flows. Specifically, from the GPS-based dataset we observe 5336 positive outflows between Tuscan mainland municipalities.

Representativeness and coverage of the study population are crucial issues when using Big Data. This is because of the specific nature of their source, in which units are not selected through a probability sampling design and generally do not cover the whole population. However, as shown by Rocco (2019), the bias due to self-selection (analogously to the bias due to non-response mechanism in survey based on probabilistic sampling) is connected to the relation between the target variable and the selection/response mechanism and it becomes negligible when this correlation is small. Following the results of Pappalardo *et al.* (2013) on the analysis of the mobility in Pisa, which showed that the mobility measured using the sample of cars with GPS is coherent with the mobility registered for all the vehicles in the municipality through the traffic sensor spread around the city, here we assume that the car commuting flows between Tuscan municipalities are low correlated with the probability of having a GPS-device in each municipality (the GPS penetration municipal rate). Therefore, the GPS sample can be considered as a representative sample for the overall population of car circulating in Tuscany.

3 A Zero-Inflated gravity model for OD commuting flows

The modelling of OD commuting flows is commonly addressed through the use and estimation of a gravity model (Anderson, 1979).

Gravity models, drawing on analogy with Newton's Law of Gravitation, typically rely on three types of factors to explain OD flows: first, origin-specific variables that characterize the ability of origin locations to generate flows; second, destination-specific variables that represent the attraction of destination locations; and, third, a function of origin-destination distance.

Standard formulation of such models assumes a multiplicative relationship between the OD frequencies and the three types of effects above defined. By taking the logarithm of both sides, a gravity model is usually defined as

$$\log(y_{ij}) = \alpha + \mathbf{X}_i^o \beta_o + \mathbf{X}_j^d \beta_d + \gamma \log(D_{ij}) + \varepsilon_{ij} \quad (1)$$

where y_{ij} denotes the flow from origin region i towards destination region j , \mathbf{X}_o and \mathbf{X}_d are explanatory variables that represent origin (o) and destination (d) characteristics, β^o and β^d are the corresponding parameters, γ reflects the effect of the distance D_{ij} between origin i and destination j and ε_{ij} is the disturbance term which

is assumed to be normally distributed, $\varepsilon_{ij} \sim N(0, \sigma^2)$. Considering n regions, the model analyse all flows for $N = n^2$ origin-destination pairs of regions (the $n \times n$ flow matrix, which contains intra-regional flows in its main diagonal and inter-regional flows in its off-diagonal elements, is vectorised by stacking the columns to form the $N \times 1$ vector of flows contained in \mathbf{y}).

When we are interested in modelling some spatial interaction counts, Flowerdew and Aitkin (1982) proposed the use of a gravity model based on the Poisson distribution instead of a model based on the Normal distribution, as in model (1). This choice seems appropriate in our context of analysis, since OD commuting flows are, by definition, non-negative counts. However, the usual Poisson regression model is not able to manage data that are characterised by overdispersion and/or an excess of zero values, and our study variable is affected by both issues.

One way to handle overdispersion is to assume a Negative Binomial distribution of the target variable, that is to generalise the Poisson distribution with an additional dispersion parameter in order to allow the conditional variance to exceed the conditional mean.

Moreover, additional overdispersion can be caused by an excess of zeros in the study variable, therefore it is important to separate the excess zeros from regular overdispersion, because the two forms of overdispersion are likely generated from different underlying processes. Hurdle and Zero-Inflated models are the most commonly applied forms of zero-augmented models and both assume that there is an additional unknown process generating excess zeros. Zero-Inflated models assume that there are two possible sources for zero-flows (true zeros and excess zeros), whereas in Hurdle models all zero values are generated by a separate process (Cameron & Trivedi, 2013).

Following these considerations, we define a Zero-Inflated Negative Binomial gravity model to investigate the commuting private car outflows y_{ij} between the Tuscan mainland municipalities:

$$y_{ij} \sim ZINB(\pi_{ij}, \mu_{ij}, \theta_{ij}) = \begin{cases} 0 & \text{with probability } \pi_{ij} \\ NB(\mu_{ij}, \theta_{ij}) & \text{with probability } 1 - \pi_{ij} \end{cases}$$

$$E(y_{ij}) = \mu_{ij}(1 - \pi_{ij})$$

$$V(y_{ij}) = \mu_{ij}(1 - \pi_{ij})(1 + \mu_{ij}(\pi_{ij} + 1/\theta_{ij}))$$

Zero-inflation model:

$$\text{logit}(\pi_{ij}) = \alpha + \beta_o x_i^o + \beta_d x_j^d + \mathbf{X}_{ij}^{od} \beta_{od} + \delta \log(D_{ij}) + v_i^o + v_j^d \quad (2)$$

Conditional models:

$$\log(\mu_{ij}) = \alpha^* + \beta_o^* x_i^o + \beta_d^* x_j^d + \mathbf{X}_{ij}^{od} \beta_{od}^* + \delta^* \log(D_{ij}) + u_i^o + u_j^d \quad (3)$$

$$\log(\theta_{ij}) = \alpha' + \delta' \log(D_{ij}) \quad (4)$$

where we assume that all three parameters of the ZINB distribution are function the driving distance D_{ij} between origin i and destination j ; moreover, we include origin- and destination-specific random effects in both the zero-inflation model (2) and the

conditional mean model (3) to capture the correlation between flows originating from the same municipality i and the correlation between the flows reaching the same municipality j

$$\begin{aligned} v_i^o &\sim N(0, \sigma_{v_o}^2), & u_i^o &\sim N(0, \sigma_{u_o}^2), \\ v_j^d &\sim N(0, \sigma_{v_d}^2), & u_j^d &\sim N(0, \sigma_{u_d}^2). \end{aligned}$$

In addition, we select the following explanatory variables to be included in model (2) and model (3):

- the working-age population of the origin municipality x_i^o (origin-specific variable), to measure the propensity of municipality i to generate commuting flows;
- the total number of employees of the destination municipality x_j^d (destination-specific variable), to measure the propensity of municipality j to attract commuting flows;
- three origin-destination-specific variables (matrix \mathbf{X}_{ij}^{od}) to measure additional characteristics which could influence the interaction between municipalities i and j : a) the indicator of origin i and destination j being in the same Local Labor Market; b) the ratio between car travel time and public transport travel time required to reach destination j from origin i ; c) the road connection quality, measured by the driving time per 100km, between origin i and destination j .

A key aspect for the estimation of a gravity model is the availability of up-to-date quality explanatory variables, particularly on OD connection times and distances. In our analysis the auxiliary variables are derived from sources of different nature: population and employees data come from ISTAT statistical datasets (the Demographic Statistics and the Statistical Register of Local Units of Active Enterprises ASIA-UL), the vehicle fleet for each Tuscan municipality is derived from the administrative dataset of the Motor Vehicle Register; and finally, the OD times and distances by means of transportation are derived from Google Maps computing services.

The estimated parameters for models (2)-(4) are presented in Table 1 and show that all the explanatory variables considered in our gravity model are statistically significant predictors and that their effect on commuting flows is as expected by common knowledge. In particular, we observe that a greater distance between two municipalities is associated with a higher probability of observing a null flow between them (estimated by model (2) and with a smaller amount of non-null commuting movements (estimated by model (3)). Analogously, if two municipalities are connected by a faster road network (that is, with a lower driving time) we expect to have larger commuting flows and a lower probability of a null flow.

The origin- and destination-specific variables of the gravity model rightfully control for the characteristics of the origin and destination locations: bigger municipalities (that is, with a larger working-age population) have a higher propensity to generate larger commuting flows whereas larger commuting flows are directed toward areas that are characterised by a higher number of employees. Moreover, the four random effects components capture the unobserved characteristics that accompanate

Table 1 Model estimates

Parameters	Zero-inflation model (2)	Mean model (3)	Dispersion model (4)
Intercept	−5.958 **	0.803 ***	0.070
Driving distance (log)	3.456 ***	−2.852 ***	0.245 ***
Driving time ($h/100km$)	2.461 ***	−0.567 ***	
Relative time (car vs public tr.)	−1.480 **	0.532 ***	
LLM indicator	−0.025 ***	0.047 ***	
Working-age population (x1000)	−0.028 ***	0.027 ***	
Number of employees (x1000)	−0.025 ***	0.047 ***	
$\sigma_{v_o}^2$	1.049 ***		
$\sigma_{v_d}^2$	0.524 ***		
$\sigma_{u_o}^2$		0.899 ***	
$\sigma_{u_d}^2$		1.145 ***	

Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ° $p < 0.1$

the flows originating from the same municipality (v^o and u^o) and the flows reaching the same municipality (v^d and u^d).

The fact that our findings confirm common knowledge supports the thesis that the commuting patterns of GPS devices' owners is similar to the commuting flows of all private car owners. If, as it seems, this assumption holds, we could use our estimated gravity model to predict the detailed private car commuting flows OD matrix between all municipalities within Tuscany.

4 Conclusion

In this paper we investigate to what extent big data collected from GPS-enabled devices, installed on private vehicles for insurance purposes, could be combined with data coming from smart data sources, like Google computing services, and from more traditional sources, like administrative and statistical datasets, in order to provide a support in producing estimates of systematic commuting flows.

This is a work in progress and many are the open questions that we intend to investigate further on.

The first step will be to assume a even more complex gravity model which, in addition to the zero-inflated structure, will be able to take into account the possi-

ble origin- and destination-based spatial autocorrelation between flows (LeSage and Pace, 2008; Sellner *et al.*, 2013).

Another relevant research topic is how to extent our analysis and findings in two direction: to estimate non-systematic flows and to estimate public transport mobility patterns, possibly by considering the different propensity on public transport usage which we expect to observe with respect to the different OD pairs.

Finally, the knowledge of some individual characteristics of the GPS devices' owners could allow us to better evaluate the representativeness of the GPS self-selected sample. However, due to the characteristics of big data sources in general and of our data source in particular, issues on privacy policy and ownership of the data could limit the availability of such knowledge. To overcome the problem, it could be necessary to conduct ad-hoc surveys or to find additional data sources.

Acknowledgements The authors acknowledge the financial support provided by the "Dipartimenti Eccellenti 2018-2022" ministerial funds.

References

1. Anderson, J.E.: A Theoretical Foundation for the Gravity Model. *American Economic Review*, **69**, 106–116 (1979)
2. Bergamaschi, S., Carlini, E., Ceci, M., Furlotti, B., Giannotti, F., Malerba, D., Mezzanzanica, M., Monreale, A., Pasi, G., Pedreschi, D., Perego, R., Ruggieri, S.: Big data research in Italy: A perspective. *Engineering*, **22**, 163–170 (2016)
3. Cameron, A., Trivedi, P.: *Regression Analysis of Count Data*, Cambridge University Press, Cambridge (2013)
4. Flowerdew, R., Aitkin, M.: A method of fitting the gravity model based on the poisson distribution. *Journal of Regional Science*, **22**, 191–202 (1982)
5. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., Trasarti, R.: Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The International Journal on Very Large Data Bases*, **20**, 695–719 (2011)
6. LeSage J.P., Pace R.K.: Spatial econometric modeling of origin-destination flows. *Journal of Regional Science*, **48**, 941–967 (2008)
7. Nanni, M., Trasarti, R., Monreale, A., Grossi, V., Pedreschi, D.: Driving profiles computation and monitoring for car insurance CRM. *ACM Trans. Intell. Syst. Technol.*, **8** (2016) doi: 10.1145/2912148
8. Pappalardo, L., Rinzivillo, S., Qu, Z., Pedreschi, D., Giannotti, F.: Understanding the patterns of car travel. *The European Physical Journal Special Topics*, **215**, 61–73 (2013)
9. Rocco, E.: Indicators for monitoring the survey data quality when non-response or a convenience sample occurs. In: Petrucci, A., Racioppi, F., Verde, R. (eds.) *New Statistical Developments in Data Science*, Springer (2019)
10. Sellner R., Fischer M.M., Koch M.: A spatial autoregressive Poisson gravity model. *Geographical Analysis*, **45**, 180–201 (2013)

Statistical calibration of the digital twin of a connected health object

Inversione statistica dei parametri di ingresso per il gemello digitale di un oggetto sanitario collegato

Nicolas Bousquet and Walid Dabachine

Abstract This article briefly describes the preparation of a clinical study on a connected health object dedicated to facilitate the prevention of lymphedema in patients treated for breast cancer (after axillary surgery), and to alert them of its manifestation in order to prevent its aggravation. A simulator of the entire physical process, or digital twin, has been developed for this purpose. Statistical calibration of its input parameters, by noisy inversion and using functional sensitivity studies, makes possible to demonstrate the feasibility of the objective by establishing one or more measurement protocols to capture the signal on a mobile object (phone or tablet) and to detect signal breaks that are physically significant. .

Abstract Questo articolo descrive brevemente la preparazione di uno studio clinico su un oggetto sanitario collegato dedicato a facilitare la prevenzione del linfedema nelle pazienti trattate per il cancro al seno (dopo la chirurgia ascellare), e ad avvertirle della sua manifestazione per prevenirne l'aggravamento. A questo scopo stato sviluppato un simulatore dell'intero processo fisico, o gemello digitale. La calibrazione statistica dei parametri di ingresso, mediante inversione rumorosa e studi di sensibilit  funzionale, consente di dimostrare la fattibilit  dell'obiettivo stabilendo uno o pi  protocolli di misura per catturare il segnale su un oggetto mobile (telefono o tablet) e rilevare interruzioni di segnale fisicamente significative.

Key words: Connected object, design of experiment, digital twin, Monte Carlo, inversion, sensitivity analysis

Nicolas Bousquet
Sorbonne Universit , 5 place Jussieu 75005 Paris & Quantmetry, 52 rue d'Anjou, 75008 Paris, France e-mail: nbousquet@quantmetry.com

Walid Dabachine
Quantmetry, 52 rue d'Anjou, 75008 Paris, France e-mail: wdabachine@quantmetry.com

1 Context

Lymphedema is a swelling of the arm caused by an accumulation of lymph, which can occur following the removal of the axillary lymph nodes during the axillary cleaning operation following a breast removal operation (see Figure 1). It can appear on the breast side a few weeks after axillary cleaning or months or even several years later. For this reason, it is necessary to remain vigilant in order to report the first symptoms as soon as possible. More generally, lymphedema is a pathology due to a deficiency of the lymphatic system. It is an organ swelling, usually of a limb due to an accumulation of lymphatic fluid between the connective tissues. There are two types of lymphedema: the mayor is congenital, and the secondary is most often induced by the surgical or radiosurgical treatment of cancer. This article focuses on upper limb lymphedema developed after axillary cleaning during breast cancer treatment. The frequency of lymphedema is 15 to 28% after axillary curing and 2.5% to 6.9% after a sentinel node procedure (selective sentinel lymphadenectomy); the frequency may increase to 44% when the surgical procedure is combined with radiotherapy. There are different risk factors: the number of lymph nodes removed, external radiotherapy, obesity, mastectomy, different physical practices. Obesity is an aggravating factor with a high prevalence: 75% of patients with lymphedema have a BMI greater than 40 kg/m^2 .



Fig. 1 Example of lymphedema.



Fig. 2 Layout of lymphometer sensors

Women with lymphedema have a degraded quality of life, with significant physical and psychological consequences. Clinically, lymphedema results in significant swelling of the limb, which reduces mobility and can cause pain in the patient. In 20 to 40% of lymphedema, infectious complications (erysipelas type, cellulitis) appear, resulting in fever, general malaise, severe pain and sometimes inflammation. In the advanced stages, sclerosis appears, causing skin lesions and alterations of ligaments and tendons. For women with dyed skin, lymphedema is a source of psychological and social difficulties: disruption of body image, disruption of identity landmarks, loss of self-esteem, anxiety (up to depressive episodes). As swelling of the limb leads to a loss of mobility and dexterity, lymphedema is an obstacle to normal professional activity, requiring professional reclassification or even permanent

work stoppage.

Today, the postoperative diagnosis of arm lymphedema in a breast cancer patient is initially made by the patient's personal presumption, who notices a clear swelling of her upper limb. Secondly, it is the medical opinion that confirms or invalidates the presumption of lymphedema of the upper limb, following a consultation. The diagnosis is therefore not instantaneous; there may be significant latency between the onset of lymphedema, suspicion of lymphedema, and final diagnosis.

2 The connected object and its digital twin

2.1 *The connected object: the lymphometer*

In 2017-2018, following a long-term collaboration with the Senology Department of the University Hospitals of Strasbourg (HUS), [5, 6, 2], a medical device was created by Quantmetry to measure the evolution of the diameter of an arm and detect swelling, the first tests of which proved promising 1. it is a sleeve that automatically takes measurements of a patient's arm circumference at 4 different locations (see Figures 2 to 4):

- 10 cm above the wrist;
- 10 cm below the elbow;
- 10 cm above the elbow;
- 20 cm above the elbow,

and thus to generate a time series of dimension 4.



Fig. 3 Physical prototype of lymphometer



Fig. 4 Lymphometer being carried on an arm.

The data is then sent to the doctor on an HDS-certified cloud server. They are aggregated and compiled and then transmitted to the patient and the doctor who follows her personally, as well as to the scientific community anonymously (if the patient agrees). The stored data is protected and secured, and the anonymity of patients is thus preserved.

This connected sleeve project addresses an important health problem: today there is currently no means of detection other than a significant swelling of the arm, visible by oneself or regular visits to the doctor who compares the circumference of the two limbs accurately.

2.2 Why a digital twin?

However, the viability of the lymphometer and its application depends, before any application for certification, on the success of a clinical study whose relevance has already been demonstrated by the HUS. The precision required by such a study requires, in particular, the ability to determine a clear measurement protocol and the acceptance of the device by the patients concerned. In practice, they will be able to operate a switch producing the measurements at different times of the day, for a certain period of time. This protocol is defined according to several unknowns:

- the time step δ_t for processing the measurement (daily, hourly?);
- the number N of times the switch is activated and its distribution by time step (multiple times repeated in a short period of time, or distributed along a time step), in order to sufficiently smooth the signal;
- the reasonable time length T the patient can wear the sleeve;
- controllable measurement noises, which can be reduced by improving fasteners sensors integrated into the sleeve.

In addition, this protocol must be robust (in the sense that it leads to a usable signal) to the following two hazards:

- the physiological stimulus that caused the arm to swell is still poorly understood; it can be sudden or progressive. It is therefore important to study a wide range of possibilities accepted by the medical profession, and verified by HUS doctors;
- Imponderable measurement noises (due to arm movements), considered random.

The variety of measured arm swelling trajectories is illustrated in Figure 5.

The work consists in preparing the clinical study of the connected object by using a digital simulator, which reproduces a large number of arm and sleeve configurations, and also simulates the production of the signal transmitted to the mobile application. This simulator is intended to test a very large number of simultaneous protocol configurations, stimuli and measurement noise and to produce one or more robust experimental protocols, the relevance of which will then be confirmed by the clinical study. A comparative illustration of the actual and measured swelling with arm noise is provided in Figure 6.

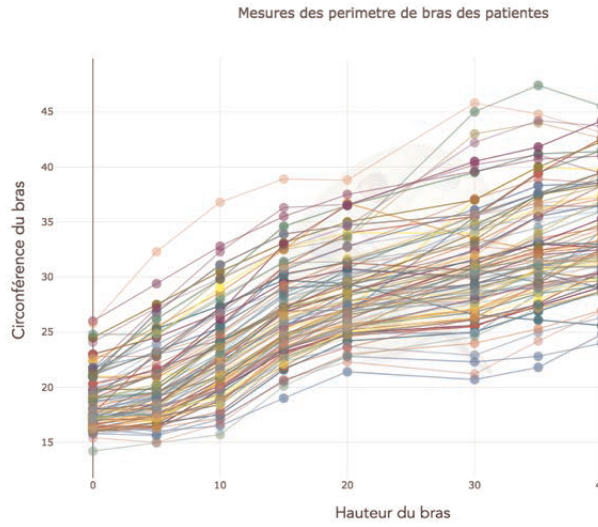


Fig. 5 Illustrations of measured swelling of an arm.

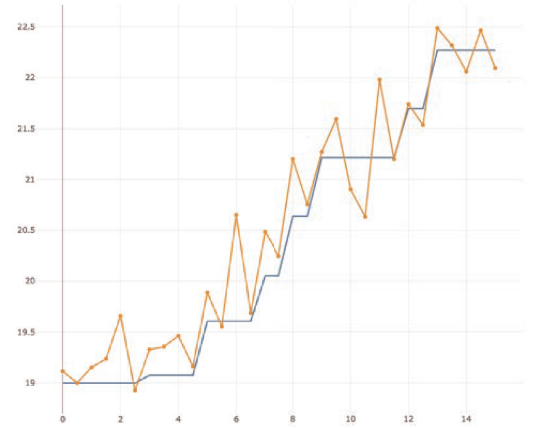


Fig. 6 Illustration of a real swelling of the arm and its measurement over time produced by the simulator.

3 Statistical calibration and sensitivity analysis

In a stationary environment, the simulation model can be formalized as

$$y(t) = g_t(X_1, X_2, X_3)$$

where g_t is the digital simulator and X_i is a set of associated parameters:

- to the experimental protocol, for $i = 1$;
- to the imponderable measurement noise, for $i = 2$;
- to the physiological stimulus (input), for $i = 3$;

The research work, which is essentially based on a statistical and probabilistic approach and will be summarized during the talk, considers the following questions:

1. Modeling a wide variety of relevant and plausible X_3 physiological stimuli, based on the expertise of HUS physicians.
2. From experimental measurements produced in the laboratory (existing and future data sets), calibrating controllable and unpredictable noise.
3. Validating the simulator by checking the agreement of the simulations obtained with experimental measurements, subject to precise knowledge of a swelling reproduced in the laboratory. In particular, it will be possible to rely on the so-called Verification, Validation and Uncertainty Quantification (VVUQ) methodologies from the scientific literature [7].

4. Conducting sensitivity analyses to prioritize sources of uncertainty, according to the most recent developments in these methodologies applied to the treatment of time series [4, 3, 1].
5. Selecting a strategy of non-parametric statistical tests capable of identifying signal breaks measured in a manner consistent with physiological stimuli, and deduce constraints on the nature of detectable stimuli.

References

1. A. Alexandarian, P.A. Gremaud, and R.C. Smith. Variance-based sensitivity analysis for time-dependent processes. *arXiv:1711.08030*, 2015.
2. A.J. Carin, S. Molière, and V. et al. Gabriele. Relevance of breast mri in determining the size and focality of invasive breast cancer treated by mastectomy: a prospective study. *World J Surg Oncol*, 15, 2017.
3. M. De Lozzo and A. Marrel. Sensitivity analysis with dependence and variance-based measures for spatio-temporal numerical simulations. *HAL-01253686*, 2016.
4. B. Iooss and P. Lemaître. A review on global sensitivity analysis methods. *In: Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, 2015.
5. M. Lodi, L. Scheer, and N. et al. Reix. Breast cancer in elderly women and altered clinico-pathological characteristics: a systematic review. *Breast Cancer Res Trea*, 2017.
6. C. Mathelin, J. Colin, and S et al. Molière. Impact du dépistage : une expérience française (impact of breast screening: a french experience). *Mise à jour du collège national des gynécologues et obstétriciens (CNGOF Proc)*, sous presse, 2018.
7. L. Swiler. Vvuq best practices in computational science/engineering problems with some thoughts about extensions/limits to complex systems models. *Report of Sandia Laboratories*, 2016.

Time Series Forecasting: Is there a role for neural networks?

Le Reti Neurali nella Previsione di Serie Storiche

Giuseppe Bruno, Sabina Marchetti, Juri Marcucci, Diana Nicoletti

Abstract Shallow and Deep artificial neural networks (NNs) have achieved top-notch performances in a wide range of different tasks such as speech and sound recognition, computer vision and natural language processing. In this paper, we employ NNs for nowcasting and forecasting some relevant macroeconomic variables. Presently linear modeling represents the mainstay for unconditional forecasting. We explore the capabilities of deep neural networks for improving upon forecast accuracy without making assumptions on the model functional form. We consider the forecast of the Index of Industrial Production (IPI) which very likely is the most widely analyzed high-frequency indicator, given the relevance of the manufacturing activity as a driver of the whole business cycle. We put forward three kind of ANN models, in different situations they outperform benchmark models at short time horizons.

Abstract *Le reti neurali, con uno o diversi strati nascosti, hanno raggiunto prestazioni eccezionali in un'ampia gamma di problematiche di stima. Fra esse ricordiamo: il riconoscimento del suono, della voce, di immagini e l'elaborazione automatica di testi. In questo articolo esse sono impiegate nell'ambito della previsione di rilevanti variabili macroeconomiche. Ad oggi la modellazione lineare costituisce il fondamento per le previsioni incondizionate. In questo lavoro esploriamo le potenzialità delle reti neurali di migliorare l'accuratezza previsiva senza formulare ipotesi sulla forma funzionale dei modelli. Consideriamo la previsione dell'indice di produzione industriale (IPI) che rappresenta uno degli indicatori ad alta frequenza più utilizzati anche per la rilevanza del settore manifatturiero come protagonista dell'intero ciclo economico. Nel lavoro proponiamo tre tipi di reti neurali, che in alcune circostanze producono errori di previsione inferiori a quelli forniti dai modelli tradizionali.*

JEL classification: C22, C53, C45

Keywords: Deep Learning, Forecasting, Time Series, Industrial Production Index

G. Bruno, S. Marchetti, J. Marcucci and D. Nicoletti
Bank of Italy, Via Nazionale 91, e-mail: giuseppe.bruno@bancaditalia.it

1 Introduction

One of the main tasks of policy-makers and analysts is that of regularly assessing the state of the economy. This is usually realized by extracting reliable signals from high frequency indicators and by using them to provide the policy-makers with early pictures of the economic situation in the short-run. In this paper we explore the role of neural networks to provide a model-free framework for macroeconomic time-series. These models have been originally inspired by the biological neural structure of the human brain functioning. The foundations of what has now known as *neural computing*, were laid in 1943 by McCulloch and Pitts who defined the threshold behaviour of a single neuron [16]. Later, Rosenblatt described a probabilistic binary classifier based on an embryonic form of neural network [17]. After more than fifteen years of very scarce contributions, we had a revival of the subject for the impressive progress achieved by the computing technology. Early in 1972 Grossberg introduced the learning concept of neural network in response to arbitrary classes of patterns; later in the same year, derived neural networks from psychological postulates about punishment and avoidance [11, 12]. From the 1980s there has been a steady growing enthusiasm for neural network applications. Between 1982 and 1984, Hopfield published two papers describing physical system which could emulate the parallel computing capabilities of the brain [14]. The following years have witnessed a flood of theoretical and empirical contributions in many different tasks ranging from image recognitions to natural language processing.

In this paper we mainly address three research questions: i) Do we have margins for improving forecasting accuracy in macroeconomics modelings?, ii) Can we easily pin down some neural network architecture enabling time series forecasting?, and iii) Are the available software tools capable to abstract away the computational and programming complexity? We address these questions by analyzing the Italian Industrial Production Index (IPI). IPI is acknowledged as a crucial macroeconomic indicator to forecast short-term evolution of the Gross Domestic Product of many industrialized countries, given the relevance of the manufacturing activity as the driver of the whole business cycle [5]. The index is produced monthly, yet its release is characterized by a significant publication delay that limits its inclusion in forecasting. Nevertheless, it attracts lots of interest as research topic to a central bank willing to achieve good *nowcasts* and to enable thus better policy decisions. Traditional econometric models were proposed to forecast the behavior of IPI, ranging from ARMA, ARIMA to VAR and BVAR [1], and to Bridge models [3, 10].

In this work we test the nowcasting ability of some kind of neural networks against benchmark linear models.

2 Methods

Starting from their prototypical single neuron form in the 1950s, neural networks (NNs) were abandoned for over 20 years because of limitations in available com-

puting resources. They were increasingly applied in recent years, thanks to more data availability and much greater computing power, to produce breakthroughs in the accuracy of many Machine Learning classification tasks; see [8] for a gentle introduction.

A NN is an oriented computational graph whose nodes, called neurons, are processing units, while edges describe the connection pattern among the nodes. The computational graph of a NN is customarily described by the following building blocks:

1. An input layer, with m neurons,
2. An output layer, with n neurons (we assume $n = 1$ throughout the present paper),
3. A set of K Hidden layers, $K \geq 1$.

This general architecture has a finer relevant distinction between *shallow* ($K = 1$) and *deep* ($K > 1$) networks. Each layer is characterised by a cardinality, that is by the number of *neurons* it contains. The processing task of each neuron consists in two simple steps. The first one is a linear combinations among its input, the second one computes a function defining the final state of the neuron. For a fixed architecture, each neuron receives, processes and forwards information accordingly. One of the main reason of the wide popularity of NNs lies in the theoretical property of shallow ones to be universal approximators of any continuous function of several variables on a compact set [7]. These results play an important role in determining boundaries of efficacy of the considered networks.

Our focus was on two classes of fully connected architectures: *Feed Forward* (FNN) and *Recurrent* (RNN, [18]) networks.

FNNs are the most basic form of NN. A FNN is composed of one or more Hidden

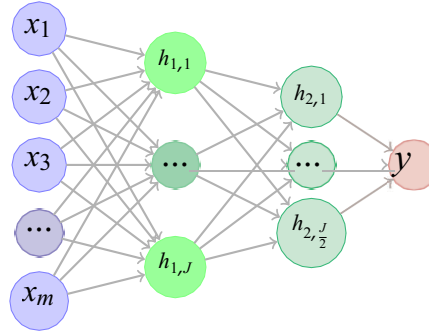


Fig. 1 A Feed-Forward neural network with two hidden layers. The network has an Input layer with m neurons, two Hidden layers with pyramid structure, and a single-node Output layer y .

layers of fully connected neurons. Each node in a layer gets inputs from every other belonging to the previous one, combines and transforms them via a fixed activation function, until it finally passes the information to every neuron in the successive layer. We restricted our search to FNNs with $K = 2$ and pyramid structure [15]. An

example of FNN is shown in Fig. 1, where $n = 1$. Due to the scarce number of observations, our search was limited to those structures with at most N/m neurons in the first Hidden layer, N and m being the number, respectively, of observations in the training set and input variables [19]. Issues related to the insufficient number of observations are further discussed below.

RNNs are networks *with memory*. These architectures originate from sequence-based research where there is temporal/order dependency. Fig. 2 (Left panel) depicts a RNN.¹ Each Hidden layer in a RNN receives the current values of Input nodes, along with the output produced by the preceding Hidden layer in the sequence. Several types of recurrent structures were proposed in the literature, e.g. [6]. We restricted our search to shallow RNNs, with Long Short Term Memory (LSTM) cells [13, 9]. LSTM cells are equipped with Forget, Input and Output gates that confer them both long-run memory and forgetting abilities, allowing new parts of the sequence to remain relevant. Formally, each LSTM cell receives the following three inputs at every time t : the current values of Input nodes \mathbf{x}^t , its previous *cell state* c_j^{t-1} , and *hidden cell state* h_j^{t-1} , $j = 1, \dots, J$, $t = 1, \dots, T$; see Fig. 2 (Right panel).

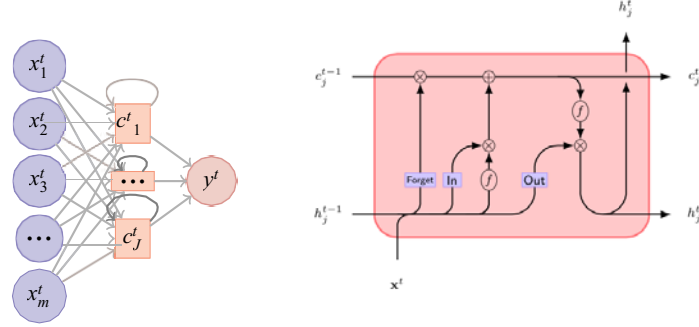


Fig. 2 Left panel: A shallow recurrent neural network, with m input variables and hidden layer with cardinality J . Right-panel: Detailed representation of a Long-Short Term Memory cell from the hidden layer of the recurrent neural network.

For any NN, we consider several *Training and Validation Process* (TVP) strategies: parameters of the network are estimated on a subset of observations, while hyper-parameters for the optimization procedure are tuned on a validation subset of data. To this purpose, we partitioned our dataset into Training, Validation and Test sets. Splitting was deterministic, based on the series' temporal order. All architectures considered recursively underwent TVP based on an expanding window. In all setups, the TVP was initialized with 70% of observations, say until t^t , $t^t = t^{(70\%)}$, \dots , $T-1$. The first 85% of the window's data points was used for training, whereas the remaining was used for validation [4]; see the upper row of Fig. 3 for an intuition.

¹ An alternative representation to Fig. 2 consists of an unfolded sequence of FNNs.

The NN was trained to nowcast target variable, say IPI^t at $t^t < t \leq T$, and the result

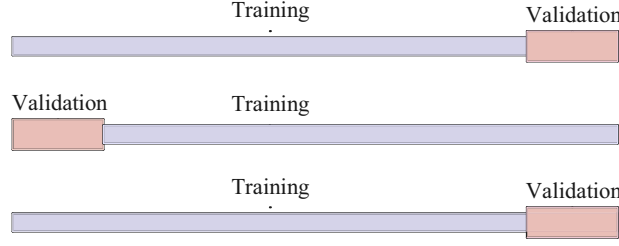


Fig. 3 Training-Validation Process. The first 85% obs. are used for training, the remaining for validation (top), the first 15% obs. are held out for validation and training is performed based on the last 85% (middle), training of the network is based on the whole set of available observations and the last 15% are used for validation (bottom).

was stored. The window would then iterate the following: i) expand by one step (from t^t to t), ii) undergo TVP adjusting the size of the subsets to leave proportions unchanged, i.e. 85% and 15%, with the parameters initialized to the estimates from the previous step, iii) forecast the next value in time for IPI, and store it. The expanding process stopped when the last forecast was produced, i.e. at $t^t = (T - 1)$. We separately accounted for $m = 12$ *soft* and $m = 9$ *non-soft*, or *hard*, indicators, based on current practice for the IPI forecasting. Particularly, details on the non-soft indicators, that correspond to the variables that are used to construct the PMI, may be found in [2].

Different scenarios for TVP are outlined in the next paragraph, and depicted in Fig. 3. Changes in the TVP had a significant impact on the nowcasting accuracy of our networks. This was largely expected, due to two different issues: the first was a consequence of the scarce stability of the model, given the small amount of data points. To partially overcome this limitation, search in the space of candidate network structures was restricted to those complying with the rule-of-thumbs mentioned above. The second issue referred to a general economic circumstance, namely the 2008 great financial crisis (GFC). For any fixed architecture, forecasts produced by the TVP outlined above would suffer the under-representation of after-GFC data points in the training set. However, tuning would eventually push the estimates in the validation sample toward the lower level established after 2008. To emphasize the role of observations following GFC in the training process, we proposed two solutions. The first proposal trivially augmented the dataset with dummy indicator I_{2008} , equal to 1 for all observations recorded until December 2018, then equal to 0. Also, we reversed the approach from [4] and used the last 85% units for training, the first 15% for validation; see middle row of Fig. 3. While this scheme proved successful in accounting for most recent information, validation would eventually push estimates toward the higher regime of GFC, before 2008. This led us to a third proposal for the TVP, based on a recursive window with overlapping sets, that we refer throughout as auto-TVP scheme. All records in the expanding window were used for training,

while validation would account for the last 15%; see bottom row of Fig. 3. This approach managed to provide quite accurate forecasts, as reported by Table 1 below. Finally, an auto-TVP rolling, or sliding, window scheme was also considered, with fixed 70% size.

Input	Model and TVP	Horizon	MAPE (Std Dev)	RMSE (Std Dev)
Hard Indicators	FNN and Expanding window	1-year	2.50% (2.67%)	3.78 (5.96)
		2-years	4.93% (4.65%)	6.75 (9.11)
	FNN and Rolling window	1-year	2.77% (3.50%)	4.56 (7.29)
		2-years	4.60% (4.51%)	6.53 (11.06)
Soft Indicators + I_{2008}	FNN and Expanding Window	1-year	1.83% (2.40%)	3.41 (8.07)
		2-years	3.23% (2.58%)	4.28 (5.02)
Soft Indicators	FNN Rolling Window	1-year	1.82% (1.42%)	2.36 (2.86)
		2-years	4.50% (3.96%)	6.18 (8.84)

Table 1 Performance evaluation of FNNs under different TVP strategies.

3 Benchmarking and comparing with Linear Models

The models we considered would take as input J lagged indicators, H real-time features, and K lagged values of IPI itself. Formally,

$$IPI^t = \sum_{j=0}^J \theta_j X_j^{t-1} + \sum_{h=0}^H \theta_h X_h^t + \sum_{k=1}^K \theta_k IPI^{t-k} + \varepsilon^t \quad (1)$$

As already mentioned, we separately accounted for $J = 12$ soft and $J = 9$ hard indicators, $H = 3$ real-time indicators, namely the northern power consumption, the amount of road and rail cargo transports, and $K = 1$ lagged values of the target variable.

The autoregressive linear model based on soft indicators yielded an overall Root Mean Squared Error (RMSE) equal to 0.03, on the period 2010-2018.

As for the second set of hard features, evidence of a structural break in the linear relationship between these indicators and IPI was previously acknowledged by the econometric literature [2]. This is more apparent when troughs of the latter occur. As a result, the linear model implemented revealed unstable, and performed poorly in forecasting, with RMSE=0.13, and Mean Absolute Percentage Error (MAPE) equal to 11.89%.

Details on the yearly results for the *non-soft* setting may be found in Table 2, where NNs with several TVP scenarios were considered. Remarkably, the linear model is outperformed by NNs in most cases. See also Table 3.

A further application accounted for all indicators as input to a linear model, successively reduced by stepwise selection process, i.e. $J = 21$. Performance of the

Time Series Forecasting: Is there a role for neural networks?

model was robust to integration of the dataset, with a slight decrease recorded by the RMSE, reaching 0.02.

MAPE	FNN-a	FNN-b	FNN-c	FNN-Dum	LSTM	FNN-Rol	Lin
2010	7.07%	2.48%	1.92%	1.73%	2.37%	3.01%	19.39%
2011	4.18%	2.99%	2.52%	1.21%	2.48%	2.17%	17.68%
2012	4.51%	7.14%	2.91%	3.01%	5.82%	2.39%	11.75%
2013	3.87%	19.09%	1.51%	1.72%	2.65%	1.76%	12.56%
2014	2.30%	20.58%	1.27%	2.38%	2.13%	1.28%	11.35%
2015	1.92%	18.39%	3.34%	2.33%	1.87%	4.08%	8.81%
2016	2.32%	16.11%	3.06%	4.52%	1.68%	5.37%	9.34%
2017	5.99%	12.23%	3.26%	2.65%	3.23%	2.65%	9.02%
2018	15.68%	8.10%	2.24%	5.57%	3.71%	0.75%	10.87%
Overall	4.02%	13.02%	2.50%	2.57%	2.87%	2.77%	11.89%

Table 2 Yearly and overall nowcasting performance of different NNs with hard indicators.

Model and TVP	Input	MAPE (Improvement on linear model)
FNN; Expanding Window	Hard Indicators	2.50% (-79.01%)
FNN; Expanding Window	Hard Indicators+ I_{2008}	2.57% (-78.42%)
FNN; Rolling Window	Hard Indicators	2.77% (-76.75%)
LSTM; Recursive Window	Hard Indicators	2.87% (-75.90%)
LSTM; Recursive Window	Hard Indicators+ I_{2008}	3.03% (-74.53%)
FNN; Recursive Window	Soft Indicators+ I_{2008}	1.90% (-9.91%)
FNN; Rolling Window	Soft Indicators	1.83% (-13.13%)

Table 3 Comparisons of the best NNs and improvement over the corresponding linear model.

4 Concluding remarks

In this paper we have employed neural network technique for macroeconomic forecasting. In particular we have applied these techniques to the IPI forecasting task. As a result, some of the model architecture provided competitive near-term forecasting performance, while not requiring any domain knowledge on the relationships and dynamics among the features considered. However, there are at least two relevant issues to emphasize. First, learning of neural network models can require a high level of computational power and take a long time to converge. As a second point, network performance depends heavily on the amount of training data, especially for deep architectures. As a consequence, hyper-parameter tuning plays a crucial role when training of a neural network is based on few observations.

There are a few ways that we expect to improve forecasting accuracy as we keep working on this project. First, we anticipate that the inclusion of a whole set of in-

dicators, with monthly or higher frequency, should sensibly improve our models. Also, techniques for data augmentation would likely favor more stable predictions, both in the training and in the hold-out sample. A second avenue we deem promising is the adoption of more advanced architectures, such as convolutional and time delay networks. Furthermore, we think that the employment of deep learning models could be integrated in our models by taking advantage more extensively of parallel processing paradigms with open or commercial software.

References

1. Gianni Amisano and Massimiliano Serati. Bvar models and forecasting: a quarterly model for the emu-11. *Statistica*, 1:51–70, 2002.
2. Valentina Aprigliano. The relationship between the pmi and the italian index of industrial production and the impact of the latest economic crisis. *Bank of Italy Temi di Discussione (Working Paper) No.*, 820, 2011.
3. Alberto Baffigi, Roberto Golinelli, and Giuseppe Parigi. Bridge models to forecast the euro area gdp. *International Journal of Forecasting*, 20:447–460, 2004.
4. Filippo Maria Bianchi, Enrico Maiorino, Michael C Kampffmeyer, Antonello Rizzi, and Robert Jenssen. *Recurrent Neural Networks for Short-Term Load Forecasting: An Overview and Comparative Analysis*. Springer, 2017.
5. Guido Bulligan, Roberto Golinelli, and Giuseppe Parigi. Forecasting monthly industrial production in real-time: From single equations to factor-based models. *Empirical Economics*, 39:303–336, 10 2009.
6. Holk Cruse. *Neural networks as cybernetic systems*. Thieme Stuttgart, 1996.
7. George Cybenko. Approximation by superposition of sigmoidal functions. *Mathematical of Control, Signal and Systems*, 2:303–314, 1989.
8. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.
9. Felix Gers, Nicol Schraudolph, and Schmidhuber Juergen. Learning precise timing with lstm recurrent networks. *Journal of Machine Learning Research*, 2:115–143, 2002.
10. Roberto Golinelli and Giuseppe Parigi. The use of monthly indicators to forecast quarterly gdp in the short run: An application to the g7 countries. *International Journal of Forecasting*, 26:77–94, 2007.
11. Stephen Grossberg. Neural expectation: cerebellar and retinal analogs of cells fired by learnable or unlearned pattern classes. *Kybernetik*, 10:49–57, 1972.
12. Stephen Grossberg. A neural theory of punishment and avoidance. *Mathematical Biosciences*, 15:39–67, 1972.
13. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
14. John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Biophysics*, 81:3088–3092, 1984.
15. Timothy Masters. *Practical Neural Network Recipes in C++*. Morgan Kaufmann, 1993.
16. Warren McCulloch and Walter Pitts. A logical calculus immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.
17. Frank Rosenblatt. The perceptron, a perceiving and recognizing automaton (project para). *Cornell Aeronautical Laboratory*, 85-460, 1957.
18. David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.
19. Shuxiang Xu and Ling Chen. A novel approach for determining the optimal number of hidden layer neurons for fnns and its application in data mining. 2008.

Modelling weighted signed networks

Modellazione di reti segnate pesate

Alberto Caimo and Isabella Gollini

Abstract In this paper we introduce a new modelling approach to analyse weighted signed networks by assuming that their generative process consists of two models: the interaction model which describes the overall connectivity structure of the relations in the network without taking into account neither the weight nor the sign of the dyadic relations; and the conditional weighted signed network model describes how the edge signed weights form given the interaction structure. We then show how this modelling approach can facilitate the interpretation of the overall network process. Finally, we adopt a Bayesian inferential approach to illustrate the new methodology by modelling the Sampson's influence network.

Abstract *In questo articolo, introduciamo un nuovo approccio modellistico per analizzare reti segnate pesate supponendo che il loro processo generativo sia costituito da due modelli: il modello d'interazione che descrive la struttura connettiva generale della rete senza tenere in considerazione né il peso né il segno degli archi; e il modello per la rete segnata pesata che descrive come gli archi segnati pesati si creino condizionatamente alla struttura d'interazione. Mostriamo quindi come questo approccio modellistico possa facilitare l'interpretazione del processo relazionale generale. Per concludere, adottiamo un approccio inferenziale bayesiano per illustrare la nuova metodologia attraverso la modellazione della rete di Sampson sulle relazioni d'influenza.*

Key words: Signed networks, weighted networks, exponential-family network models, Bayesian inference.

Alberto Caimo
Technological University Dublin, Ireland. e-mail: alberto.caimo@dit.ie

Isabella Gollini
University College Dublin, Ireland. e-mail: isabella.gollini@ucd.ie

1 Exponential random graph models

Relations between actors on social networks often consist of positive and negative interactions. And typically these interactions have weights assigned to them. The relational structure of a network graph is represented by an adjacency matrix \mathbf{y} whose elements $y_{i,j}$ are defined by a value corresponding to the intensity of the interaction between any pair of nodes.

Exponential random graph models (ERGMs) are a particular class of discrete linear exponential families [8, 17] which represent the probability distribution of a network \mathbf{Y} on a fixed set of nodes as:

$$p(\mathbf{y}|\boldsymbol{\theta}) = h(\mathbf{y}) \exp\{\boldsymbol{\theta}'s(\mathbf{y}) - \psi(\boldsymbol{\theta})\}, \quad (1)$$

where $h(\mathbf{y})$ is a reference distribution [10] specifying the model for the data before any network effect is considered; $s(\mathbf{y})$ is a known vector of p network statistics measuring the quantity of some selected sub-graph configurations in the network [14], $\boldsymbol{\theta} \in \mathbb{R}^p$ is the parameter vector associated to the vector of network statistics, and $\psi(\boldsymbol{\theta})$ is a normalising constant which is typically computationally difficult to evaluate for all but trivially small networks [12]. The dependence hypothesis at the basis of the ERGMs is that the observed network structure is the result of a generative process in which edges self organise into sub-network configurations. There is a wide range of possible network configurations which gives the flexibility to adapt ERGMs to various different contexts. A positive parameter value for θ_i results in a tendency for the certain configuration corresponding to $s_i(\mathbf{y})$ to be observed in the data than would otherwise be expected by chance.

2 ERGMs for weighted signed networks

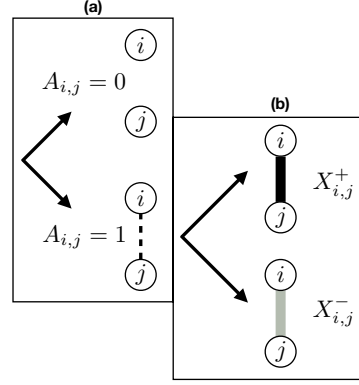
A weighted signed network graph between N nodes can be described by $N \times N$ adjacency matrix \mathbf{Y} where:

$$Y_{i,j} = \begin{cases} y_{i,j} \neq 0, & i \text{ connected to } j; \\ y_{i,j} = 0, & i \text{ not connected to } j. \end{cases}$$

The connection value of $y_{i,j}$ represents the weight of positive and negative edges between nodes.

ERGMs have recently been generalised to binary signed networks [9]. In this paper we adopt a new modelling approach for weighted signed networks by assuming the existence of two distinct processes: the interaction process determining the presence or absence of an interaction between the nodes (see also [11]), and the conditional weighted signed process which is describing the joint structure of the positive and negative weight relations given the interaction process. We distinguish between an interaction variable \mathbf{A} and a weighted signed variable \mathbf{X} by assuming that the prob-

Fig. 1 Structure of the model:
(a) interaction process; **(b)**
 conditional weighted signed
 network process.



ability of Y represents the joint probability of A and X :

$$\Pr(Y = y) = \Pr(X = x, A = a) = \Pr(X = x | A = a) \times \Pr(A = a).$$

In particular, we assume that the overall weighted signed network process can be jointly modelled by a joint conditional weighted ERGM process for the positive weighted edges (X^+) and the negative weighted edges (X^-), so that for any dyad (i, j) we have:

$$\begin{cases} \Pr(Y_{i,j} = y_{i,j} | Y_{i,j} > 0) = \Pr(X_{i,j}^+ = x_{i,j}^+ | A_{i,j} = 1, \theta_X^+) \times \Pr(A_{i,j} = 1 | \theta_A); \\ \Pr(Y_{i,j} = y_{i,j} | Y_{i,j} < 0) = \Pr(X_{i,j}^- = x_{i,j}^- | A_{i,j} = 1, \theta_X^-) \times \Pr(A_{i,j} = 1 | \theta_A), \end{cases}$$

where:

$$A_{i,j} = \begin{cases} a_{i,j} = 1, & i \text{ connected to } j; \\ a_{i,j} = 0, & i \text{ not connected to } j. \end{cases}$$

It is important to notice that positive and negative processes are not conditionally independent given A as the two signed structures X^+ and X^- are mutually exclusive given A .

We therefore propose to model the interaction process assuming that $A | \theta_A \sim \text{ERGM}(\theta_A)$ and the weighted network process assuming that X is modelled by two joint weighted signed ERGM processes: one for positive edge relations and one for negative edge relations with parameters θ_X^+ and θ_X^- , respectively. Figure 1 shows the structure of the interaction/weighted signed modelling framework proposed.

The conditional weighted ERGM processes can be defined according to specific forms of weighted network model. ERGM modelling approaches for weighted networks include the multi-valued curved ERGMs [18], generalised ERGMs for inference on networks with continuous edge values [6]; Geometric/Poisson reference ERGMs for ordinal/count networks [10]; and the hierarchical multilayer ERGM approach for polytomous networks [4].

3 Bayesian inference

Bayesian methods are becoming increasingly popular as techniques for modelling social networks. Following the Bayesian paradigm, prior distribution is assigned to θ . The posterior distribution of θ given the observed network data \mathbf{y} is:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}.$$

From an ERGM viewpoint, direct evaluation of $p(\theta|\mathbf{y})$ requires the calculation of both the likelihood $p(\mathbf{y}|\theta)$ and the marginal likelihood or model evidence $p(\mathbf{y})$ which are typically intractable. According to our modelling framework, the parameter posterior distribution defined in Section 2 can be written as:

$$p(\theta_X^+, \theta_X^-, \theta_A | \mathbf{x}^+, \mathbf{x}^-, \mathbf{a}) \propto p(\mathbf{x}^+, \mathbf{x}^- | \mathbf{a}, \theta_X^+, \theta_X^-) p(\theta_X^+, \theta_X^-) p(\mathbf{a} | \theta_A) p(\theta_A). \quad (2)$$

where:

- $p(\mathbf{a} | \theta_A) \propto h(\mathbf{a}) \exp\{\theta_A s(\mathbf{a})\}$ is the interaction ERGM likelihood;
- $p(\mathbf{x}^+, \mathbf{x}^- | \mathbf{a}, \theta_X^+, \theta_X^-)$ is the joint weighted signed ERGM likelihood conditional on the interaction relations;
- $p(\theta_X^+, \theta_X^-)$ and $p(\theta_A)$ are the prior parameter distributions.

To estimate the parameter posterior density defined in Equation 2, we adapt the approximate exchange algorithm for Bayesian ERGMs [1, 2] implemented in the Bergm package [3] for R [13].

4 Application

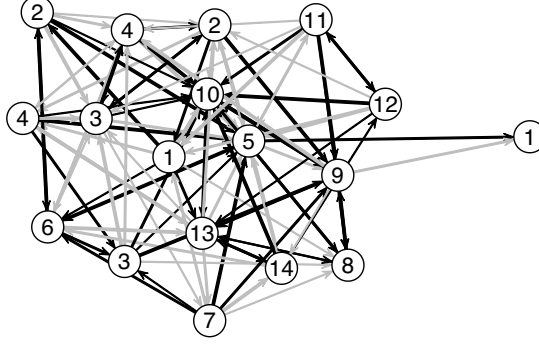
Sampson's monk directed network [15] contains ratings between monks related to a crisis in a cloister in New England (USA). In particular, we focus on the positive/negative influence between monks and we want to use the ERGM generative process defined above to describe the connectivity structure of the weighted signed directed network.

4.1 Model specification

We include the following network statistics for the binary undirected interaction ERGM model:

- Edge statistic (`edges`) is the number of edges in the network capturing the network density effect.

Fig. 2 Graph structure of the Sampson’s monk network. Weighted black edges correspond to positive influence relations; gray edges correspond to negative influence relations.



- Geometrically weighted edgewise shared partner statistic (*gwesp*) captures the tendency towards transitivity, i.e., the tendency of edges to be connected through multiple triadic relations simultaneously [16]. We fix the decay parameter of the *gwesp* statistic to be equal to 2.

We assume that the conditional weighted signed network model follows a constrained ERGM process with a uniform/truncated geometric reference distribution [10, 7] so that $h(\mathbf{x}^+, \mathbf{x}^-; \mathbf{a}) = 1$.

We include the following directed network statistics for the joint conditional weighted signed ERGM model:

- Weighted-sum statistic (*sum*) is the sum of the edge values capturing the weighted density effect.
- Mutual-min statistic (*mutual (min)*) is the sum of the minimum weighted mutual edge value capturing the weighted network reciprocity effect.
- Transitive-weights statistic (*transitiveweights*) captures the tendency towards transitive clustering in the weighted network.
- Cyclical-weights statistic (*cyclicalweights*) captures the tendency towards cyclical clustering in the weighted network.

It is important to emphasise that the set of the network statistics to include for describing the positive and negative structures is not necessarily the same so that we can formulate different connectivity hypothesis for the positive and negative weighted network processes.

4.2 Prior specification

We specify vague prior distribution for all the parameters in both interaction and conditional weighted signed network model:

$$\theta_X^+ \sim \mathcal{N}_p(\mu = 0, \Sigma = 10 \mathbf{I}_p); \theta_X^- \sim \mathcal{N}_q(\mu = 0, \Sigma = 10 \mathbf{I}_q); \theta_A \sim \mathcal{N}_r(\mu = 0, \Sigma = 10 \mathbf{I}_r).$$

where \mathbf{I} is the identity matrix. In this example, the model specification for the joint weighted signed process consist of the same set of network statistics for both the positive and the negative weighted structure.

4.3 Results

The posterior density estimates displayed in Table 1 show that the interaction network process $\mathbf{A}|\theta_A$ is sparse (negative value for the `edges` parameter) and the interactions tend to organise into triadic clusters (positive value for `gwesp(2)`). The positive `mutual(min)` parameter estimate in the conditional positive weighted signed ERGM process $\mathbf{X}^+|\mathbf{X}^-, \mathbf{A}, \theta_X^+$ explains the tendency towards reciprocation of positive influence relations. It is important to notice that the positive tendency towards clustering that we generally expect in positive weighted networks is mainly captured by the `gwesp(2)` effect in the interaction model. The negative triadic parameter (`transitiveweights` and `cyclicalweights`) estimates for the conditional negative weighted signed ERGM process $\mathbf{X}^-|\mathbf{X}^+, \mathbf{A}, \theta_X^-$ confirm the assumption of the structural balance hypothesis of “the enemy of an enemy is a friend” or “the friend of an enemy is an enemy” [5].

Table 1 Parameter posterior estimates for the interaction / weighted signed network model.

	Network statistic	Mean	Std. Err.
Interaction	<code>edges</code>	−1.42	0.51
	<code>gwesp(2)</code>	0.19	0.06
Positive weights	<code>sum</code>	−0.11	0.21
	<code>mutual(min)</code>	0.82	0.38
	<code>transitiveweights</code>	0.03	0.20
	<code>cyclicalweights</code>	0.01	0.16
Negative weights	<code>sum</code>	0.37	0.21
	<code>mutual(min)</code>	0.22	0.40
	<code>transitiveweights</code>	−1.56	0.35
	<code>cyclicalweights</code>	−0.64	0.17

4.4 Model assessment

A way to examine the fit of the data to the estimated posterior distribution of the parameters is to implement a graphical Bayesian goodness-of-fit procedure. In the Bayesian context, simulated networks are simulated from a sample of 100 parameter values randomly drawn from the estimated posterior distribution and compared to the observed data according to some network statistics. The plots in Figure 3

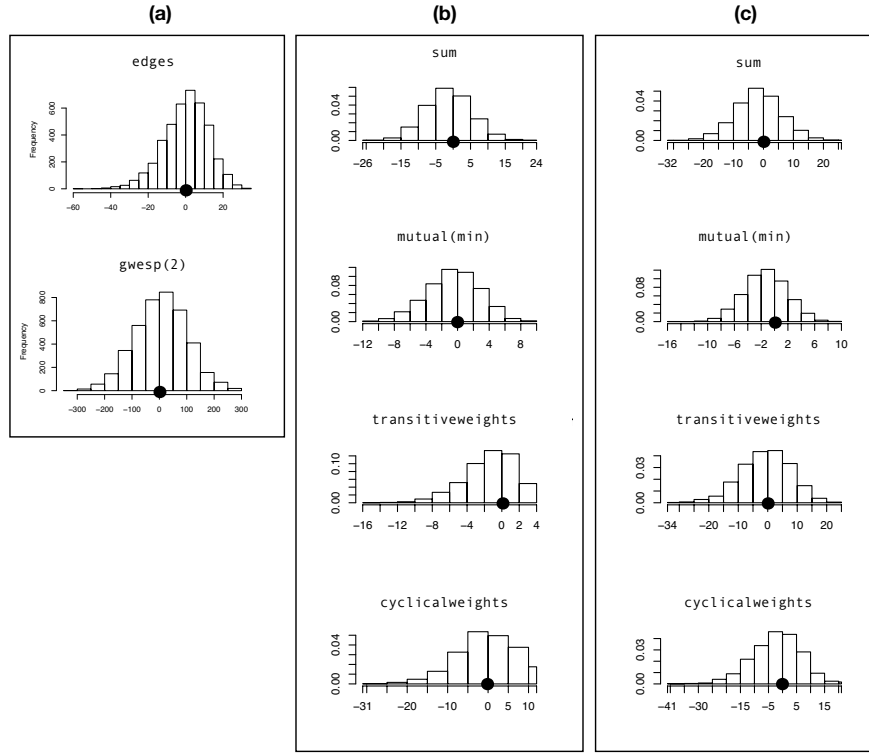


Fig. 3 Goodness of fit diagnostics plots showing the distribution of a set of networks simulated from the estimated posterior density of Table 1 centred on the observed network statistics for (a) Interaction model; (b) Positive weighted signed model; (c) Negative weighted signed model.

suggest that the proposed model is a reasonable fit to the observed data as most of the probability mass of the simulated network statistics is concentrated around the observed network statistics.

5 Discussion

We have presented a flexible model able to capture the dependence structure of weighted signed networks. In particular, given the interaction between nodes in the network we have proposed to model the weighted signed network dependencies by introducing a weighted signed ERGM processes for joint modelling the structure of negative and positive edges and adopting a Bayesian approach. As demonstrated in the illustration, the model is able to facilitate the interpretation of the complex dependence structure of weighted signed networks by making use of interpretable network effects and the assessment of structural balance in signed networks.

References

1. Caimo, A., Friel, N.: Bayesian inference for exponential random graph models. *Social Networks* **33**(1), 41 – 55 (2011)
2. Caimo, A., Friel, N.: Bayesian model selection for exponential random graph models. *Social Networks* **35**(1), 11 – 24 (2013)
3. Caimo, A., Friel, N.: Bergm: Bayesian exponential random graphs in R. *Journal of Statistical Software* **61**(2), 1–25 (2014). URL <http://www.jstatsoft.org/v61/i02/>
4. Caimo, A., Gollini, I.: A multilayer exponential random graph modelling approach for weighted networks. arXiv preprint arXiv:1811.07025 (2018)
5. Cartwright, D., Harary, F.: Structural balance: a generalization of heider’s theory. *Psychological review* **63**(5), 277 (1956)
6. Desmarais, B.A., Cranmer, S.J.: Statistical inference for valued-edge networks: The generalized exponential random graph model. *PloS one* **7**(1), e30,136 (2012)
7. Garlaschelli, D.: The weighted random graph model. *New Journal of Physics* **11**(7), 073,005 (2009)
8. Holland, P.W., Leinhardt, S.: An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association* **76**, 33–65 (1981)
9. Huitsing, G., Van Duijn, M.A., Snijders, T.A., Wang, P., Sainio, M., Salmivalli, C., Veenstra, R.: Univariate and multivariate models of positive and negative networks: Liking, disliking, and bully–victim relationships. *Social Networks* **34**(4), 645–657 (2012)
10. Krivitsky, P.N.: Exponential-family random graph models for valued networks. *Electronic journal of statistics* **6**, 1100 (2012)
11. Lerner, J.: Structural balance in signed networks: Separating the probability to interact from the tendency to fight. *Social Networks* **45**, 66–77 (2016)
12. Morris, M., Handcock, M.S., Hunter, D.R.: Specification of exponential-family random graph models: terms and computational aspects. *Journal of statistical software* **24**(4), 15–48 (2008)
13. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2011). URL <http://www.R-project.org>
14. Robins, G., Pattison, P., Kalish, Y., Lusher, D.: An introduction to exponential random graph models for social networks. *Social Networks* **29**(2), 169–348 (2007)
15. Sampson, S.F.: A novitiate in a period of change. an experimental and case study of social relationships. Ph.D. thesis, Cornell University (1968)
16. Snijders, T.A.B., Pattison, P.E., Robins, G.L., S., H.M.: New specifications for exponential random graph models. *Sociological Methodology* **36**, 99–153 (2006)
17. Strauss, D., Ikeda, M.: Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association* **5**, 204–212 (1990)
18. Wyatt, D., Choudhury, T., Bilmes, J.A.: Discovering long range properties of social networks with multi-valued time-inhomogeneous models. In: AAAI (2010)

Issues on Bayesian nonparametric measures of disclosure risk

Questioni su misure Bayesiane nonparametriche di rischio di "disclosure"

Federico Camerlenghi, Cinzia Carota and Stefano Favaro

Abstract Consider a microdata sample from a finite population, such that each record contains two disjoint types of information: identifying and sensitive information. Any decision about releasing data is supported by the estimation of measures of disclosure risk, which are functionals of the number of records with a unique combination of values of identifying variables. The work of [7] first explored the use of exchangeable random partitions to estimate a common measure of disclosure risk: the number of unique sample records that are also unique population records. In this paper we revisit the work of [7] from a Bayesian nonparametric perspective, and we discuss new potential research directions in the fields.

Abstract *Si consideri un campione di microdati da una popolazione finita, dove ogni record contiene due informazioni disgiunte: informazioni identificative e sensibili. Ogni decisione sul rilascio dei dati è supportata dalla stima di misure del rischio di "disclosure", che sono funzione del numero di record con una combinazione unica di valori delle variabili identificative. Il lavoro di [7] ha studiato per la prima volta l'uso delle partizioni aleatorie scambiabili per la stima di un comune rischio di "disclosure": il numero di record campionari unici che sono anche record di popolazione unici. In questo articolo rivisitiamo il lavoro di [7] in ottica Bayesiana nonparametrica, e discutiamo nuove direzioni di ricerca.*

Key words: Bayesian nonparametrics; Dirichlet process prior; disclosure risk; empirical Bayes; exchangeable random partitions; identifying and sensitive information.

Federico Camerlenghi

University Milano - Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy. e-mail: federico.camerlenghi@unimib.it

Cinzia Carota

University of Torino, Lungo Dora Siena 100 A, 10153 Torino, Italy. e-mail: cinzia.carota@unito.it

Stefano Favaro

University of Torino, Corso Unione Sovietica 218bis, 10134 Torino, Italy. e-mail: stefano.favaro@unito.it

1 Introduction

Consider a microdata sample (X_1, \dots, X_n) from a finite population of size $N > n$, such that each sample record X_i contains two disjoint types of information: identifying information and sensitive information. Identifying information consists of the values of categorical variables, which might be matchable to known units of the population. See, e.g., [1] and [9] for a comprehensive account on these types of information. A risk of disclosure arises from the possibility that an intruder might succeed in identifying a microdata unit through such a matching and hence be able to disclose sensitive information of the unit. To prevent disclosure, any decision about releasing data is supported by the estimation of measures of disclosure risk, which are suitable functionals of the number of records with a unique combination of values of identifying variables. Indeed, assuming no errors in the matching process or data sources, for unique records the match is guaranteed to be correct. When sample records are cross-classified according to the identifying variables, the microdata sample is partitioned in $K_n \leq n$ non-empty cells, labelled by $\{X_1^*, \dots, X_{K_n}^*\}$, $M_{i,n}$ of which have frequency i , for $i = 1, \dots, n$. Two common measures of disclosure risk are: i) the number v_1 of unique population records; ii) the number τ_1 of unique sample records that are also unique population records.

The work of [7] first explored the use of exchangeable random partition to estimate τ_1 . His ideas can be described in terms of an urn scheme, where records belonging to the cell X_i^* are depicted as balls of the same color. Using Samuels' terminology, let $(X_i)_{i \geq 1}$ be a superpopulation of colored balls belonging to a (ideally) infinite number of colors $(X_i^*)_{i \geq 1}$ with unknown composition $P = (p_i)_{i \geq 1}$, i.e. p_i is the probability of drawing a ball of color X_i^* , with $\sum_{i \geq 1} p_i = 1$ almost surely. Then, consider a population (X_1, \dots, X_N) which a random sample from P such that: i) (X_1, \dots, X_n) is an initial observable random sample from P ; ii) $(X_{n+1}, \dots, X_{N-n})$ is an additional unobservable random sample from P . In particular, (X_1, \dots, X_N) takes on the interpretation of the population records, of which the subsample (X_1, \dots, X_n) are the observable sample records. A natural way to make inference on the composition of the population (X_1, \dots, X_N) from (X_1, \dots, X_n) is to imagine sampling the remainder of the population $(X_{n+1}, \dots, X_{N-n})$ from the posterior distribution of P given (X_1, \dots, X_n) . The problem of estimating τ_1 can thus be stated as follows: given a urn whose initial composition is (X_1, \dots, X_n) , and given $(N - n)$ draws from the urn, how many colors which are unique in the initial state will remain unique at the final state?

[7] addressed this problem under an urn scheme introduced by [5] in (mathematical) population genetics. Specifically, consider an urn that initially contains only a black ball with mass $\theta > 0$, and apply iteratively the following sampling scheme: i) if we pick the black ball then it is returned with a ball of a new color with mass 1; ii) if we pick a non-black ball then it is returned with a ball of the same color with mass 1. Let $M_{i,n}$ denote the number of colors with frequency $i > 1$ after n draws, namely the number of unique sample records. By relying on the sole sampling scheme, [7] showed that the expected τ_1 is

$$m_{1,n} \frac{n + \theta - 1}{N + \theta - 1}, \quad (1)$$

with $m_{1,n}$ being the observed number of unique cells. [7] then specified θ via maximum likelihood, that is by choosing the value $\hat{\theta}$ that maximizes the likelihood function for sample records. From [5] this corresponds to solve, with respect to θ , the equation $k_n = \sum_{1 \leq j \leq n-1} \theta / (\theta + j)$, with k_n being the observed number of distinct cells. Experiments in [7] show that the estimator (1) leads to a systematic underestimation of τ_1 , which worsen as N and n become large.

In this paper we revisit and discuss the work of [7] from a Bayesian nonparametric perspective. We derive Bayesian nonparametric estimators of v_1 and τ_1 under a Dirichlet process prior ([4]) for the unknown composition P of the superpopulation. These estimators have simple closed-form expressions, and they are obtained by a direct application of results by [3] on conditional formule for Gibbs-type exchangeable random partitions. Not surprisingly, our estimator of τ_1 coincides with Samuels' estimator (1), showing that (1) is a Bayesian nonparametric estimator of τ_1 , with respect to a squared loss function, under a Dirichlet process prior for P . This provides with a Bayesian derivation of Samuels' estimator, and it paves the way to discuss the following issues: i) the problem of uncertainty quantification for Samuels' estimator; ii) the interpretation of (1) as a Bayesian smoothed version of the nonparametric estimator $m_{1,n}n/N$, which is the naive estimator of τ_1 when $m_{1,n}$ is used as an estimator of v_1 ([1]); iii) the problem of estimating v_1 and τ_1 under generalized Dirichlet priors, e.g., the two parameter Poisson-Dirichlet prior ([6]) and, more generally, any prior in the class of Poisson-Kingman models ([6]); iv) the problem of extending the approach of [7] to deal with the presence of structurally empty cells (structural zeros).

2 Bayesian nonparametric measures of disclosure risk

For any $\theta > 0$, let $(v_i)_{i \geq 1}$ be a collection of independent random variables with v_i distributed according to a Beta distribution with parameter $(1, \theta)$, and let $(p_i)_{i \geq 1}$ be a sequence of random variables defined as follows: $p_1 = v_1$ and $p_i = v_i \prod_{1 \leq j \leq i-1} (1 - v_j)$, for any $i \geq 2$. Furthermore, let $(X_i^*)_{i \geq 1}$ be a collection of random variables independent of $(v_i)_{i \geq 1}$ and independent and identically distributed according to a nonatomic probability measure α_0 . The Dirichlet process prior with parameter θ and base distribution α_0 is defined as the law of the random probability measure $P_{\theta, \alpha_0} = \sum_{i \geq 1} p_i \delta_{X_i^*}$. We assume a Dirichlet process prior on the unknown composition P of the superpopulation. Then, the sample record (X_1, \dots, X_n) is a random samples from P_{θ, α_0} , i.e.,

$$\begin{aligned} X_1, \dots, X_n | P_{\theta} &\stackrel{iid}{\sim} P_{\theta, \alpha_0} \\ P_{\theta, \alpha_0} &\sim \mathcal{D} \end{aligned} \quad (2)$$

with \mathcal{D} denoting the law of P_{θ, α_0} . Due to the (almost sure) discreteness of \mathcal{D} , we expect ties in a sample (X_1, \dots, X_n) from P_{θ, α_0} . Specifically, (X_1, \dots, X_n) features $K_n = k_n \leq n$ distinct types, labelled by $\{X_1^*, \dots, X_{K_n}^*\}$, with frequencies $\mathbf{N}_n = (N_{1,n}, \dots, N_{K_n,n}) = (n_{1,n}, \dots, n_{k_n,n})$ such that $N_{i,n} \geq 1$ and $\sum_{1 \leq i \leq K_n} N_{i,n} = n$. The distribution of the random variable $(K_n, N_{1,n}, \dots, N_{K_n,n})$ models the composition of the (observable) sample records. We refer to [6] for a comprehensive account on the distribution of $(K_n, N_{1,n}, \dots, N_{K_n,n})$, which is known as the Ewens sampling formula.

We now present a description of the composition of the additional (unobservable) random sample $(X_{n+1}, \dots, X_{N-n})$. As recalled in the introduction, this is assumed to be a random sample from the posterior distribution of P_θ given (X_1, \dots, X_n) . In particular, due to the conjugacy of the Dirichlet process prior ([4]), the law of P_θ given (X_1, \dots, X_n) is the law of the Dirichlet process $P_{\theta+n, \alpha_0+e_n}$, with $e_n = n^{-1} \sum_{1 \leq i \leq n} \delta_{X_i}$. Let $\{Y_1^*, \dots, Y_{J_{N-n}}^*\}$ be the labels of the $J_{N-n} \leq N-n$ distinct types in the additional sample $(X_{n+1}, \dots, X_{N-n})$ that do not coincide with any of the X_i^* . Moreover, let $0 \leq V_{N-n} \leq n$ be the number of X_{n+i} 's that do not coincide with any of the X_i^* 's, and set

- i) $\mathbf{S}_{N-n} = (S_{1,N-n}, \dots, S_{K_n,N-n})$, where $S_{j,N-n}$ is the number of X_{n+i} 's that coincide with the label X_j^* , for any $j = 1, \dots, K_n$, such that $S_{j,n} \geq 0$ and $\sum_{1 \leq j \leq K_n} S_{j,n} = n - V_n$;
- ii) $\mathbf{R}_{N-n} = (R_{1,N-n}, \dots, R_{J_{N-n},N-n})$, where $R_{j,N-n}$ be the number of X_{n+i} 's that coincide with the label Y_j^* , for any $j = 1, \dots, J_{N-n}$, such that $R_{j,n} \geq 1$ and $\sum_{1 \leq j \leq J_n} R_{j,n} = V_n$.

The conditional distribution of the random variable $(\mathbf{S}_{N-n}, V_{N-n}, J_{N-n}, \mathbf{R}_{N-n})$ given $(K_n, N_{1,n}, \dots, N_{K_n,n})$ models the composition of the additional unobservable sample. We refer to [3] for additional details on $(\mathbf{S}_{N-n}, V_{N-n}, J_{N-n}, \mathbf{R}_{N-n})$, as well as for its numerous distributional properties, i.e. conditional Ewens sampling formula.

Under the Bayesian nonparametric model (2), we consider the problem of estimating ν_1 and τ_1 . First, we give a formal definition of ν_1 and τ_1 in terms of the random variables $K_n, N_{i,n}, J_{N-n}, S_{i,N-n}$ and $R_{i,N-n}$ introduced above. In particular, we can write

$$\nu_1 = \sum_{i=1}^{K_n} \mathbb{1}_{\{N_{i,n}+S_{i,N-n}=1\}} + \sum_{i=1}^{J_n} \mathbb{1}_{\{R_{i,N-n}=1\}}$$

and

$$\tau_1 = \sum_{i=1}^{K_n} \mathbb{1}_{\{N_{i,n}+S_{i,N-n}=1\}}.$$

We consider the problem of deriving a Bayesian nonparametric estimator of ν_1 and τ_1 with respect a squared loss function. That is, we want to compute the conditional expectation of ν_1 given the sample records (X_1, \dots, X_n) , and the conditional expectation of τ_1 given the sample records (X_1, \dots, X_n) . These results follows by a direct application of Equation 23 and Equation 24 in [3]. It is worth pointing out that results in [3] can be applied also to derive high-order moments of the posterior distribution. Furthermore, these moments may lead, with additional effort, to the

posterior distribution of v_1 given (X_1, \dots, X_n) . Here, for the sake of simplicity, we focus on the expected value of the posterior distribution.

Proposition 1. Let (X_1, \dots, X_n) be sample records consisting of $K_n = k_n$ non-empty cells with corresponding frequencies $\mathbf{N}_n = (n_{1,n}, \dots, n_{k_n,n})$ and such that $m_{1,n} = \sum_{1 \leq i \leq k_n} \mathbb{1}_{\{n_{i,n}=1\}}$. Under the Bayesian nonparametric model (2), one has

$$\mathbb{E} \left[\sum_{i=1}^{K_n} \mathbb{1}_{\{N_{i,n} + S_{i,N-n}=1\}} \mid X_1, \dots, X_n \right] = \frac{\theta + n - 1}{\theta + N - 1} m_{1,n}$$

and

$$\mathbb{E} \left[\sum_{i=1}^{J_n} \mathbb{1}_{\{R_{i,N-n}=1\}} \mid X_1, \dots, X_n \right] = \frac{\theta}{\theta + N - 1} (N - n).$$

Proposition 1 provides a simple closed-form expression for a Bayesian nonparametric estimator, with respect to a squared loss function, of v_1 and τ_1 . In particular, we have

$$\hat{v}_1 := \mathbb{E}[v_1 \mid X_1, \dots, X_n] = \frac{\theta + n - 1}{\theta + N - 1} m_{1,n} + \frac{\theta}{\theta + N - 1} (N - n) \quad (3)$$

and

$$\hat{\tau}_1 := \mathbb{E}[\tau_1 \mid X_1, \dots, X_n] = \frac{\theta + n - 1}{\theta + N - 1} m_{1,n}. \quad (4)$$

Observe that the number $M_{1,n}$ of unique sample records is sufficient to estimate the measures τ_1 and v_1 . That is $M_{1,n}$ is the sole information contained in the random sample (X_1, \dots, X_n) that is required to estimate τ_1 and v_1 . To the best of our knowledge, the estimator \hat{v}_1 in Equation (3) is the first example of a Bayesian nonparametric estimator of v_1 . With regards to Equation (4), the Bayesian nonparametric estimator $\hat{\tau}_1$ coincides with Samuels' estimator (1). Then, our result provides with a proper Bayesian derivation of Samuels' estimator (1) as posterior expectation under a Dirichlet process prior. While this is an interesting result, it is not surprising due to the well-known interplay between Hoppe's sampling scheme and the Dirichlet prior ([5]).

3 Discussion

We revisited the work of [7] from a Bayesian nonparametric perspective. In particular, we introduced a Bayesian nonparametric estimator of v_1 , and we showed that Samuels' estimator (1) is a Bayesian nonparametric estimator of τ_1 , with respect to a squared loss function, under a Dirichlet process prior for P . Then, the estimator $\hat{\tau}_1$ may be viewed as a Bayesian smoothed version of the naive estimator $m_{1,n}n/N$ discussed by [1] and [9]. In particular, as N and n become large, the effect of the smoothing prior parameter θ vanishes, and $\hat{\tau}_1$ becomes approximately the naive

estimator. This motivates the underestimation phenomenon discussed in [7]. Our Bayesian derivation of Samuels' estimator (1) introduces a natural way to deal with the problem of uncertainty quantification, which was missing in the original work of [7]. With some effort results in Section 3.1. of [3] can be applied to derive the posterior distribution of v_1 given the sample records (X_1, \dots, X_n) , and the posterior distribution of τ_1 given the sample records (X_1, \dots, X_n) . These posterior distributions then provide with natural tools for quantifying uncertainty for the estimators \hat{v}_1 and $\hat{\tau}_1$.

Given our derivation of Samuels' estimator (1), one may consider the use of different prior distributions for the unknown composition P of the superpopulation. For instance, one may consider the use of the celebrated two-parameter Poisson-Dirichlet prior ([6]). Under this prior assumption one may still rely on results in [3] to derive the posterior distribution of v_1 given the sample records (X_1, \dots, X_n) , and the posterior distribution of τ_1 given the sample records (X_1, \dots, X_n) . In general, results [3] provide useful tools to deal with Bayesian nonparametric estimation of v_1 and τ_1 under sufficiently flexible prior assumptions. It remains an open problem to adapt our Bayesian nonparametric approach to deal with structural zeros. In that regard a concrete direction of research would consist in making use of a Dirichlet process prior with a spike and slab base measure (see, e.g., [8] and [2]). In other terms, the nonatomic base distribution α_0 is replaced by a base distribution of the form $\alpha_0(\zeta) = \zeta \delta_0 + (1 - \zeta) \alpha_0$, with $\zeta \in [0, 1]$ and α_0 being a nonatomic distribution. The parameter ζ is then used to include the information on structural zeros, taking on the interpretation of the proportion of structural zeros in the population or records.

References

- [1] BETHLEHEM, J.G., KELLER, W.J. AND PANNEKOEK, J. (1990). Disclosure control of microdata. *J. Amer. Stat. Ass.*, **85** 38–45.
- [2] CANALE, A., LIJOI, A., NIPOTI, B. AND PRÜNSTER, I. (2017). On the Pitman-Yor process with spike and slab base measure. *Biometrika*, **104** 681–697.
- [3] FAVARO, S., LIJOI, A. AND PRÜNSTER, I. (2013). Conditional formulae for Gibb-type exchangeable random partitions. *Ann. Appl. Probab.*, **23**, 1721–1754.
- [4] FERGUSON, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.
- [5] HOPPE, F.H. (1984). Pólya-like urns and the Ewens sampling formula. *J. Math. Biol.*, **20**, 91–94.
- [6] PITMAN, J. (2006). *Combinatorial Stochastic Processes*. Ecole d'Été de Probabilités de Saint-Flour XXXII. Lecture notes in mathematics, Springer - New York.
- [7] SAMUELS, S.M. (1998). A Bayesian, species-sampling-inspired approach to the unique problem in microdata disclosure risk assessment. *J. Off. Statist.*, **14**, 373–383.
- [8] SCARPA, B. AND DUNSON, D. (2009). Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics*, **65** 772–780.
- [9] SKINNER, C.J., MARSH, C., OPENSHAW, S. AND WYMER, C. (1994). Disclosure control for census microdata. *J. Off. Statist.*, **10**, 31–51.

Hierarchies of nonparametric priors

Gerarchie di distribuzioni iniziali nonparametriche

Federico Camerlenghi, Stefano Favaro and Lorenzo Masoero

Abstract Hierarchical processes are extremely popular tools in Bayesian nonparametrics, particularly suited to induce probabilistic dependence across observations coming from different, though similar, studies. In the present paper we discuss diverse classes of hierarchical processes tailored for species and feature models. We first focus on hierarchical random probability measures [8], clarifying how these priors are useful in presence of populations of animals composed by different species. We then move to feature models, which generalize species sampling models by allowing every observation to belong to more than one species, now called features. We introduce a new class of hierarchical priors designed for such a setting, whose distributional properties will be the subject of an ongoing work [3].

Abstract *I processi gerarchici sono strumenti molto utili in Statistica Bayesiana nonparametrica, spesso indicati per indurre dipendenza probabilistica tra osservazioni che provengono da esperimenti diversi, ma simili tra loro. Nel presente lavoro discuteremo diverse classi di processi gerarchici pensati per modelli di specie o caratteristiche. Prima di tutto introdurremo le misure di probabilità aleatorie gerarchiche [8], chiarendo come queste possano essere usate nel contesto delle specie. Successivamente ci focalizzeremo su una generalizzazione dei precedenti modelli, dove ogni osservazione può appartenere a più specie contemporaneamente, che ora prendono il nome di caratteristiche. Introdurremo una nuova classe di distribuzioni iniziali per tali modelli, le cui proprietà saranno oggetto di un lavoro futuro [3].*

Federico Camerlenghi

Department of Economics, Management and Statistics, University of Milano - Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy e-mail: federico.camerlenghi@unimib.it

Stefano Favaro

Department of Economics and Statistics, University of Torino, Corso Unione Sovietica 218/bis, 10134 Torino, Italy e-mail: stefano.favaro@unito.it

Lorenzo Masoero

Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA e-mail: lom@mit.edu

Key words: Hierarchical processes, species models, feature models, completely random measures, partial exchangeability

1 Introduction

In the present paper we consider observations coming from multiple, though similar, populations, and we define general classes of nonparametric priors suitable to carry out full Bayesian inference under the framework of partially exchangeable data. To fix the notation, we consider a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and we denote by $(\mathbb{Z}, \mathcal{Z})$ a Polish space equipped with the corresponding Borel σ -algebra; we further denote by $\mathbb{P}_{\mathbb{Z}}$ the space of all probability measures over $(\mathbb{Z}, \mathcal{Z})$. The j th observation of population i will be indicated as $Z_{i,j}$, and it is assumed to be a \mathbb{Z} -valued random element defined on $(\Omega, \mathcal{A}, \mathbb{P})$. The ideally infinite sequences of observations $\mathbf{Z}_i := (Z_{i,j})_{j \geq 1}$, for $i = 1, \dots, d$, are supposed to be partially exchangeable [11], more precisely this means that $(\mathbf{Z}_1, \dots, \mathbf{Z}_d) \stackrel{d}{=} (\pi_1 \mathbf{Z}_1, \dots, \pi_d \mathbf{Z}_d)$, where $\pi_i \mathbf{Z}_i = (Z_{i, \pi_i(j)})_{j \geq 1}$ and π_1, \dots, π_d are finite permutations on the natural numbers. Thanks to the de Finetti representation theorem, the sequences \mathbf{Z}_i 's are partially exchangeable iff there exists a probability measure Q_d over $\mathbb{P}_{\mathbb{Z}}^d$, such that

$$\mathbb{P} \left[\bigcap_{i=1}^d \bigcap_{j=1}^{n_i} (Z_{i,j} \in A_{i,j}) \right] = \int_{\mathbb{P}_{\mathbb{Z}}^d} \prod_{i=1}^d \prod_{j=1}^{n_i} p_i(A_{i,j}) Q_d(dp_1, \dots, dp_d) \quad (1)$$

for any $(n_1, \dots, n_d) \in \mathbb{N}^d$ and for any collection of Borel sets $A_{i,j} \in \mathcal{Z}$, as $j = 1, \dots, n_i$, $i = 1, \dots, d$. The measure Q_d in (1) is termed *de Finetti measure* and it works as a prior distribution to carry out Bayesian inference. A large amount of Bayesian nonparametric literature has recently focused on the choice of the distribution Q_d to induce dependence across the d groups of observations. In the present paper we will define classes of priors which induce probabilistic dependence in a hierarchical fashion. More precisely we discuss two constructions particularly suited for species and feature models.

In the species setting each observation $Z_{i,j}$ represents the species' label of an animal belonging to a population composed by different species, whose proportions are supposed to be completely unknown. Species estimation finds numerous applications in several scientific disciplines and its importance has grown in recent years. See, e.g., [6, 7]. General classes of nonparametric priors for this type of observations have been recently introduced and investigated by [8], where the authors provide suitable generalizations of the well-known Hierarchical Dirichlet Process [18, 19] along with fundamental distributional results. The model proposed in [8] will be discussed and reviewed in Section 3.

Feature models generalize species sampling models by allowing every observation to belong to more than one species, which are now called features. Each observation $Z_{i,j}$ is endowed with a finite set of features which are chosen among a possi-

bly infinite collection and may be conveniently represented as counting measures $Z_{i,j} = \sum_{k \geq 1} a_{i,j,k} \delta_{x_k}$, where the x_k 's represent the features, while the $a_{i,j,k}$'s are the associated weights. These models have been first applied in ecology to indicate the presence or absence of animals in a trap [10], in such a situation the features represent species' labels, while $a_{i,j,k} = 0$ if the trap contains species k , 0 otherwise. Moreover many data analysis problems can be seen as discovering a set of latent features within a population. For example, we might be interested in learning latent topics from a set of documents or learning latent ancestral groups from genetic data [2, 16]. A Bayesian nonparametric (BNP) approach further enables the total number of latent features present in the data to be unknown and to grow as more data are collected. To achieve the full flexibility of this approach while allowing practical inference, a number of authors have developed a very general BNP formalism that allows conjugacy and thereby efficient inference for the infinite models of BNP [5, 9, 13]. Particular examples that fit within this framework have found wide application [1, 2, 4, 12, 14, 21]. However, more complex modeling framework is desired when one is provided with multiple populations, such as multiple collections of documents. For this reason in Section 4 we will define a new Bayesian hierarchical model tailored for feature models, this in turn leads us to the specification of a distribution Q_d when the observations $Z_{i,j}$ are counting measures. Posterior analysis and distribution theory for the prior presented in Section 4 will be the subject of an ongoing work by the same authors [3].

The rest of the paper proceeds as follows. The hierarchical priors we are going to introduce here are constructed by means of Completely Random Measures (CRMs), which are briefly recalled in Section 2. We then move to define hierarchical random probability measures for multiple populations of species in Section 3, we finally introduce the new class of priors designed for feature models in Section 4.

2 Basics on Completely random measures

Completely Random Measures (CRMs) are effective tools to define nonparametric priors in Bayesian literature. In the present section we recall some basics on CRMs. Let us denote by $M_{\mathbb{Z}}$ the space of all measures on $(\mathbb{Z}, \mathcal{Z})$ which are finite on bounded sets, i.e. $\mu \in M_{\mathbb{Z}}$ iff it satisfies $\mu(A) < +\infty$ for any for any bounded set $A \in \mathcal{Z}$. We further assume that $M_{\mathbb{Z}}$ is equipped with the corresponding Borel σ -algebra $\mathcal{M}_{\mathbb{Z}}$.

Definition 1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, a Completely Random Measure (CRM) $\tilde{\mu}$ is a measurable map defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in $(M_{\mathbb{Z}}, \mathcal{M}_{\mathbb{Z}})$, such that the random variables $\tilde{\mu}(A_1), \dots, \tilde{\mu}(A_k)$ are independent for any choice of disjoint Borel sets $A_1, \dots, A_k \in \mathcal{Z}$ and for any $k \geq 1$.

CRMs have been first introduced by Kingman [15], who proved that $\tilde{\mu}$ may be decomposed into the sum of three main parts: a diffuse measure, an infinite sum of random jumps at fixed locations, an infinite sum of random jumps at random

locations. In the present paper we consider CRMs of the latter type, for which there exists a sequence of positive jumps $(J_k)_{k \geq 1}$ and a sequence of random atoms $(Z_k^*)_{k \geq 1}$ such that the CRM $\tilde{\mu}$ may be written as

$$\tilde{\mu} = \sum_{k \geq 1} J_k \delta_{Z_k^*}$$

The probability distribution of a CRM may be characterized in terms of the Laplace functional, defined as

$$\mathbb{E}[e^{-\int_{\mathbb{Z}} f(z) \tilde{\mu}(dz)}] = \exp \left\{ -c \int_{\mathbb{Z}} \int_0^\infty (1 - e^{-sf(z)}) v(ds, dz) \right\} \quad (2)$$

where $v(ds, dz)$ is called the Lévy intensity of the CRM $\tilde{\mu}$ and we assume that it can be decomposed as $v(ds, dz) = \rho(s) ds cP(Dz)$, where ρ is a positive function on \mathbb{R}^+ , $c > 0$ and P is a probability measure on $(\mathbb{Z}, \mathcal{Z})$ called the base measure of the sequence. This is equivalent to assuming that the two sequences of jumps $(J_k)_{k \geq 1}$ and atoms $(Z_k^*)_{k \geq 1}$ are independent. We will write $\text{CRM}(\rho; cP)$ to denote the distribution of the CRM $\tilde{\mu}$ having the Laplace functional (2) with $v(ds, dz) = \rho(s) ds cP(dz)$.

3 Hierarchical priors for species models

In the species' setting the data point $Z_{i,j}$ represents the species' labels of individual j in the i th group of observations. In such a framework $Z_{i,j}$ comes from an unknown probability distribution, say $p_i = \sum_{k \geq 1} p_{i,k} \delta_{z_k}$, where z_k denotes the label of the species while $p_{i,k}$ is its proportion in population i . We choose a hierarchical normalized completely random measure to model our prior opinion on the vector of probabilities (p_1, \dots, p_d) . Let us first recall that we may define a random probability measure by simply normalizing a CRM $\tilde{\mu} = \sum_{k \geq 1} J_k \delta_{Z_k^*}$:

$$\tilde{p} = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{Z})} = \sum_{k \geq 1} \frac{J_k}{\bar{J}} \delta_{Z_k^*} \quad (3)$$

where $\bar{J} := \sum_{k \geq 1} J_k = \tilde{\mu}(\mathbb{Z})$. Such a construction has been proposed by [17] and (3) is called Normalized Random Measure with Independent Increments (NRMI) \tilde{p} . The distribution of \tilde{p} will be denoted as $\text{NRMI}(c, \rho; P)$. Note that \tilde{p} in (3) is well defined if one assumes that $\mathbb{P}(0 < \tilde{\mu}(\mathbb{Z}) < \infty) = 1$, which is guaranteed iff the two conditions

$$\int_0^\infty \rho(s) ds = +\infty, \quad \int_0^\infty \min\{s, 1\} \rho(s) ds < +\infty$$

are satisfied. Being provided with d different random probability measures $\tilde{p}_1, \dots, \tilde{p}_d$, one may enable dependence across them in the following hierarchical fashion:

$$\begin{aligned}\tilde{p}_i | \tilde{p}_0 &\stackrel{\text{iid}}{\sim} \text{NRMI}(c_i, \rho_i; \tilde{p}_0) \quad i = 1, \dots, d \\ \tilde{p}_0 &\sim \text{NRMI}(c_0, \rho_0; P_0).\end{aligned}\tag{4}$$

where P_0 is a diffuse measure on $(\mathbb{Z}, \mathcal{Z})$. In (4) the base measure referring to each \tilde{p}_i , for $i = 1, \dots, d$, is taken to be random and equals another NRMI \tilde{p}_0 : such a construction allows for sharing of species across $\tilde{p}_1, \dots, \tilde{p}_d$. The distribution of the vector of hierarchical NRMI's $(\tilde{p}_1, \dots, \tilde{p}_d)$ induces a prior distribution Q_d as in (1) that can be used in the context of multiple populations of species.

4 Hierarchical priors for feature models

In the feature setting we have observations from multiple populations and we are interested in constructing hierarchical feature models, for this reason we introduce a hierarchical generalization of CRMs. In such a framework a typical observation Z is a counting measure of the type $Z = \sum_{k \geq 1} a_k \delta_{x_k}$, where x_k is the feature and a_k is the weight associated with feature k . For instance Z may describe the composition of a trap, in such a situation the x_k 's are the species' labels of the animals in the whole population and $a_k \in \{0, 1\}$, with $a_k = 0$ if animal k is in the trap or 0 otherwise. A typical condition that is required is that $\sum_{k \geq 1} \mathbb{1}_{\{a_k > 0\}} < +\infty$, i.e. the total number of features associated with each observation is finite. More generally, now denote by \mathbb{X} the space of features' labels and \mathcal{X} will be the respective σ -algebra. In order to define a vector of hierarchical CRMs, first consider a base CRM $\tilde{\mu}_0$ having Lévy intensity given by $\rho_0(s)dc_0P_0(dx)$, where P_0 is a non-atomic probability measure on the Polish space $(\mathbb{X}, \mathcal{X})$, hence $\tilde{\mu}_0 \sim \text{CRM}(\rho_0; c_0P_0)$. We know that $\tilde{\mu}_0$ can be represented in terms of the points $\{(\tilde{h}_{0,k}, \tilde{x}_k)\}_{k \geq 1}$ of a marked Poisson point process on $\mathbb{R}^+ \times \mathbb{X}$ as follows

$$\tilde{\mu}_0 = \sum_{k \geq 1} \tilde{h}_{0,k} \delta_{\tilde{x}_k}, \quad \text{a.s.}$$

Conditionally given $\tilde{\mu}_0$, we now define the vector of CRMs $(\tilde{\mu}_1, \dots, \tilde{\mu}_d)$, for $d \geq 1$, as follows

$$\tilde{\mu}_i | \tilde{\mu}_0 \stackrel{d}{=} \sum_{k \geq 1} \tilde{h}_{i,k} \delta_{\tilde{x}_k} \tag{5}$$

where $(\tilde{h}_{i,k})_{k \geq 1}$ are sequences of positive jumps for population i . In order to enable dependence across groups of observations we further suppose that, given $\tilde{\mu}_0$, the $\tilde{h}_{i,k}$ are independent with distribution depending on $(\tilde{h}_{0,k}, \tilde{x}_k)$ and a parameter $c_i > 0$ which is specific to group i , more precisely

$$\tilde{h}_{i,k} | \tilde{\mu}_0 \stackrel{\text{ind}}{\sim} f_i(h | \tilde{h}_{0,k}, \tilde{x}_k, c_i)$$

where $f_i(h | \tilde{h}_{0,k}, \tilde{x}_k, c_i)$ is a density on the positive real line. In the context of BNP models, this hierarchical construction has been considered in the specific case of the beta process by [20] (see also Section 4.1 below), however no theory has been

developed so far neither for such a simple situation. Our general viewpoint will allow to find all the theoretical results, which are necessary to develop full Bayesian inference for these models.

The observations $Z_{i,j}$'s are now modeled as random counting measures, more precisely

$$Z_{i,j} | (\tilde{\mu}_1, \dots, \tilde{\mu}_d) = \sum_{k \geq 1} a_{i,j,k} \delta_{\tilde{x}_k} \quad (6)$$

for $i = 1, \dots, d$. The $(a_{i,j,k})_{i,j,k}$'s are independent random variables conditionally on the vector of random measures having the following distribution

$$\mathbb{P}(a_{i,j,k} = s | \tilde{\mu}_i) = \pi_i(\tilde{h}_{i,k}) P_i(s | \tilde{h}_{i,k}) + (1 - \pi_i(\tilde{h}_{i,k})) \delta_{\{0\}}(s),$$

this is a spike-slab specification where $a_{i,j,k} = 0$ with probability $(1 - \pi_i(\tilde{h}_{i,k}))$, otherwise $a_{i,j,k} > 0$ with probability $\pi_i(\tilde{h}_{i,k})$, finally $P_i(\cdot | \tilde{h}_{i,k})$ is a probability distribution on the positive integers and it can be interpreted as the distribution of $a_{i,j,k}$ conditional on the event $\{a_{i,j,k} > 0\}$. The sequences of observations $\{(Z_{i,j})_{j \geq 1} : i = 1, \dots, d\}$ are then partially exchangeable, and the de Finetti measure Q_d in (1) is uniquely determined by this construction. We finally suppose the validity of the following condition

$$\int_{\mathbb{X}} \int_0^{+\infty} \int_0^{+\infty} \pi_j(h) f_j(h | s, x, c_j) dh \rho(s | x) ds P_0(dx) < +\infty,$$

which is tantamount to assuming that the number of features observed in each individual is almost surely finite. Posterior analysis and the development of suitable sampling schemes for such a construction will be the subject of an ongoing work by the same authors [3].

4.1 A case studied: the hierarchical Beta–Bernoulli process

In this section we discuss a relevant example of the model (5) already introduced in [20] and termed hierarchical Indian Buffet Process. In order to do this we first define a vector of hierarchical three–parameters Beta processes $(\tilde{\mu}_1, \dots, \tilde{\mu}_d)$. In such a case the measure at the root of the hierarchy, i.e. $\tilde{\mu}_0$, is a three-parameter Beta process, namely a CRM having Lévy intensity defined by

$$\rho_0(r) dr c_0 P_0(dx) = c_0 \frac{\Gamma(1 + \alpha_0)}{\Gamma(1 - \sigma_0) \Gamma(\alpha_0 + \sigma_0)} r^{-\sigma_0 - 1} (1 - r)^{\alpha_0 + \sigma_0 - 1} dr \mathbb{1}_{(0,1)}(r) P_0(dx),$$

depending on three parameters $\alpha_0 > 0$, $c_0 > 0$ and $\sigma_0 \in (0, 1)$. Conditionally on $\tilde{\mu}_0$, the vector of CRMs $(\tilde{\mu}_1, \dots, \tilde{\mu}_d)$ is such that

$$\tilde{\mu}_i | \tilde{\mu}_0 \stackrel{d}{=} \sum_{k \geq 1} \tilde{h}_{i,k} \delta_{\tilde{x}_k}$$

where $\tilde{h}_{i,k} \in (0, 1)$ almost surely, and the $\tilde{h}_{i,k}$'s are conditionally independent, given $\tilde{\mu}_0$, with Beta distribution specified as follows

$$\tilde{h}_{i,k} | \tilde{\mu}_0 \stackrel{\text{ind}}{\sim} \text{Beta}(c_i \tilde{h}_{0,k}, c_i (1 - \tilde{h}_{0,k}))$$

in other words $f_i(\cdot | r, x, c_i)$ is the density function of a Beta random variable. The vector $(\tilde{\mu}_1, \dots, \tilde{\mu}_d)$ so defined is said to be a vector of hierarchical Beta processes. If in addition the $a_{i,j,k}$'s appearing in (6) are independent Bernoulli random variables with parameter $\tilde{h}_{i,k}$, given $\tilde{\mu}_i$, we obtain the hierarchical Indian Buffet Process of [20]. The terminology is due to the fact that $Z_{i,j}$ is a Bernoulli process and $a_{i,j,k} = 1$ if one observes feature \tilde{x}_k , or 0 otherwise.

References

1. Blei, D.M.: Probabilistic topic models. *Comm. ACM*, **55**, 77–84 (2012)
2. Blei, D.M., Ng, A.Y., and Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022 (2003)
3. Broderick, T., Camerlenghi, F., Favaro, S., Masoero, L.: Posterior representations of hierarchical completely random measures in trait allocation models. *Manuscript in preparation* (2019)
4. Broderick, T., Mackey, L., Paisley, J., Jordan, M.I.: Combinatorial clustering and the beta negative binomial process. *IEEE Trans. Pattern Anal. Mach. Intell.*, **37**, 290–306 (2015)
5. Broderick, T., Wilson, A.C., Jordan, M.I.: Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli*, **24**, 3181–3221 (2018)
6. Bunge, J., Fitzpatrick, M.: Estimating the number of species: a review. *J. Amer. Statist. Assoc.*, **88**, 364–373 (1993)
7. Bunge, J., Willis, A., Walsh, F.: Estimating the number of species in microbial diversity studies. *Annu. Rev. Stat. Appl.*, **1**, 427–445 (2014)
8. Camerlenghi, F., Lijoi, A., Orbanz, P., Prünster, I.: Distribution theory for hierarchical processes. *Ann. Statist.*, **47**, 67–92 (2019)
9. Campbell, T., Huggins, J.H., How, J.P., Broderick, T.: Truncated random measures. *arXiv preprint arXiv:1603.00861* (2016)
10. Colwell, R., Chao, A., Gotelli, N.J., Lin, S., Mao, C.X., Chazdon, R.L., Longino, J.T.: Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, **5**, 3–21 (2012)
11. de Finetti, B.: Sur la condition d’équivalence partielle. *Actualités scientifiques et industrielles*, 5–18 (1938)
12. Ghahramani, Z., Griffiths, T.L.: Infinite latent feature models and the Indian buffet process. In: *Advances in Neural Information Processing Systems*, pp. 475–482 (2006)
13. James, L.F.: Bayesian Poisson calculus for latent feature modeling via generalized Indian buffet process priors. *Ann. Statist.*, **45**, 2016–2045 (2017)
14. Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., Ueda, N.: Learning systems of concepts with an infinite relational model. In *AAAI* (2006)
15. Kingman, J.: Completely random measures. *Pacific J. Math.*, **21**, 59–78 (1967)
16. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959 (2000)
17. Regazzini, E., Lijoi, A., Prünster, I.: Distributional results for means of normalized random measures with independent increments. *Ann. Statist.*, **31**, 560–585 (2003)
18. Teh, Y.W., Jordan, M.I.: Hierarchical Bayesian nonparametric models with applications. In: *Bayesian Nonparametrics*, pp. 158–207. Cambridge University Press, Cambridge (2010)

19. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 1566–1581 (2006)
20. Thibaux, R., Jordan, M.I.: Hierarchical beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics*, pp. 564–571 (2007)
21. Zhou, M., Carin, L.: Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, **37**, 307–320 (2015)

Issues with Nonparametric Disclosure Risk Assessment

Questioni sull'Analisi Nonparametrica del Rischio di "Disclosure"

Federico Camerlenghi, Stefano Favaro, Zacharie Naulet and Francesca Panero

Abstract Whenever an agency releases microdata files containing sensitive information about a sample of individuals, the major concern must be to protect the identity of the participants from what is known as *disclosure risk*. An intruder, in fact, could match the published file with previously available information and disclose the identity of the participants. One of the approaches to tackle this type of risk is the statistical quantification of it by means of the estimation of the number of sample uniques that are also population uniques, i.e. the individuals whose risk of being disclosed is the highest. In this paper, we review the literature about parametric and nonparametric estimation of this measure and present new potential research questions.

Abstract *Ogniquale volta un'agenzia pubblica microdati contenenti informazioni sensibili riguardanti un campione di individui, la principale preoccupazione deve essere la protezione dell'identità dei partecipanti dal cosiddetto rischio di "disclosure". Un intruso, infatti, potrebbe incrociare i dati pubblicati con informazioni precedentemente disponibili e scoprire l'identità dei partecipanti. Uno degli approcci per contrastare questo tipo di rischio è la quantificazione statistica di esso, per mezzo della stima del numero di unici nel campione che sono anche unici nella popolazione, cioè quelli per cui rischio di "disclosure" è più alto. In questo lavoro*

Federico Camerlenghi

University Milano - Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy. e-mail: federico.camerlenghi@unimib.it

Stefano Favaro

University of Torino, Corso Unione Sovietica 218bis, 10134 Torino, Italy. e-mail: stefano.favaro@unito.it

Zacharie Naulet

University of Toronto, Sidney Smith Hall, 100 St George St, ON M5S 3G3 Toronto, Canada. e-mail: znaulet@utstat.toronto.edu

Francesca Panero

University of Oxford, 24-29 St Giles', OX1 3LB Oxford, United Kingdom. e-mail: francesca.panero@stats.ox.ac.uk

presentiamo la letteratura esistente su metodi parametrici e nonparametrici di stima di questa misura e offriamo nuove possibili direzioni di ricerca.

Key words: disclosure risk assessment, microdata sample, parametric inference, nonparametric inference

1 Introduction

In the study of survey and census data, microdata represent a precious source of information to perform data analysis, but to preserve their correctness and availability, sensitive information contained in them must be protected from intruders. If a sample of these data has to be published, it is necessary for the agency releasing it to understand if a participant whose information is in that sample could be disclosed. This crucial issue is known as *disclosure risk*. An intruder, in fact, could use personal or publicly available information and match them with the published sample, attempting to disclose the identity, and therefore sensitive information, of the participants.

A microdata sample $\mathbf{X}(n) = (X_1, \dots, X_n)$ is a collection of records for n subjects containing identifying information and sensitive information. The sample is supposed to come from a population \mathbf{X} of finite size $\bar{n} = n + \lambda n$, $\lambda > 0$.

One of the approaches considered to tackle disclosure risk aims at estimating some quantities related to the sample. In particular, a popular choice is to estimate the number of sample uniques that are also population uniques. These represent the individuals whose record is different from anybody else's both in the sample and in the population, and for whom the risk of being identified is the highest.

To formalize the definition, we cross-classify the sample according to potentially identifying variables in $K_n \leq n$ cells whose frequency is $(Y_1(\mathbf{X}, n), \dots, Y_{K_n}(\mathbf{X}, n))$: $Y_j(\mathbf{X}, n)$ represents the number of individuals in $\mathbf{X}(n)$ belonging to the j -th cell. Of course, $\sum_{j=1}^{K_n} Y_j(\mathbf{X}, n) = n$. We can repeat this classification with the whole population, creating $K_{\bar{n}}$ cells, each of frequency $Y_j(\mathbf{X}, \bar{n})$, $j = 1, \dots, K_{\bar{n}}$.

According to this classification, the number of population uniques that are also sample uniques is

$$\tau_1(\mathbf{X}, n, \lambda) = \sum_{j \geq 1} \mathbb{1}_{\{Y_j(\mathbf{X}, n)=1\}} \mathbb{1}_{\{Y_j(\mathbf{X}, n+\lambda n)=1\}}. \quad (1)$$

A number of different approaches have been tested to produce estimators for τ_1 . The majority of them imposed parametric or semiparametric distributional structures on the population and sample, both frequentist and Bayesian. See, e.g., [1], [15], [16], [12], [13], [18], [9], [10], [2] and [3].

In this paper we review some of these methods, highlighting the intrinsic difficulties that the estimation, in particular nonparametric, might encounter. In Section 2 we illustrate a few parametric and nonparametric estimators, while Section 3 is devoted to a numerical comparison of the presented methods. Finally, in section 4 we present issues and open problems in nonparametric estimation.

2 Existing Approaches

Among the first Bayesian parametric approaches to quantify disclosure risk, we find [1] that relies on the empirical Bayes approach of [4]. They considered a population of \bar{n} individuals divided in $K_{\bar{n}}$ species. To each species j is assigned a superpopulation parameter $p_j > 0$ and a random variable F_j representing the population frequency of that species, assumed to be distributed as a $Poisson(\bar{n}p_j)$. f_j , $j = 1, \dots, K_{\bar{n}}$, are the corresponding values in the sample. Their approach exploited the classical *Poisson-Gamma* model, in which data are distributed as Poisson random variables, whose parameters are modelled by Gamma prior distributions. More formally:

$$\begin{aligned} P_j &\stackrel{\text{iid}}{\sim} \text{gamma}(\alpha, \beta), \\ F_j | P_j = p_j &\sim \text{Poisson}(Np_j). \end{aligned}$$

We should have $\sum_{j=1}^{K_{\bar{n}}} P_j = 1$, but for the sake of simplicity the authors assumed $\mathbb{E}[\sum_{j=1}^{K_{\bar{n}}} P_j] = K_{\bar{n}}\alpha\beta = 1$. Under this model, the expected value of population uniques is $U_p = \mathbb{E}\left[\sum_{j=1}^{K_{\bar{n}}} \mathbb{1}_{F_j=1}\right] = N(1 + N\beta)^{-(1+\alpha)}$. The method of moments or maximum likelihood were used to find estimates $\hat{\alpha}, \hat{\beta}$, and therefore an estimator \hat{U}_p . As an estimator for τ_1 , they proposed the sample proportion of the estimated population uniques:

$$\hat{\tau}_1^B := \hat{U}_p \frac{n}{N}. \quad (2)$$

If $K_{\bar{n}}$ is not available, [1] suggested to estimate it assuming a uniform distribution over the cells, hence

$$\hat{K}_{\bar{n}} = \frac{\bar{n}K_n}{\sum_{j=1}^{K_n} \mathbb{1}_{\{Y_j(\mathbf{X},n)=1\}}}.$$

Since this approach was proven not to be robust, [17] proposed an improvement of this model, adding the specification of the distribution of f_j 's:

$$f_j | P_j = p_j \stackrel{\text{indep}}{\sim} \text{Poisson}(np_j),$$

Under this specification, the probability that an individual is a population unique, given that he is an observed sample unique, is

$$\mathbb{P}(\text{population unique} \mid \text{sample unique}) = \left(\frac{1+N\beta}{1+n\beta} \right)^{-(1+\alpha)}.$$

Therefore, the estimator for sample uniques that are also population uniques becomes:

$$\hat{\tau}_1^S := K_n \left(\frac{1+N\hat{\beta}}{1+n\hat{\beta}} \right)^{-(1+\hat{\alpha})}. \quad (3)$$

A naive nonparametric estimator of τ_1 is the sampling fraction, with respect to the population, of the number of sample uniques $Z_1(\mathbf{X}, n) = \sum_{j=1}^{K_n} Y_j(\mathbf{X}, n)$. This estimator was first discussed in [1] and [17], and it is defined as

$$\hat{\tau}_1^{\mathcal{N}} := Z_1(\mathbf{X}, n) \frac{n}{\bar{n}}. \quad (4)$$

[15] exploits Bayesian nonparametric ideas, and in particular a Dirichlet process prior ([6]) on the p_j 's to derive a smoothed version of the naive estimator (4). In particular, the proposed estimator was the following

$$\hat{\tau}_1^{\mathcal{D}} := Z_1(\mathbf{X}, n) \frac{n + \vartheta - 1}{\bar{n} + \vartheta - 1}, \quad (5)$$

where ϑ is the concentration parameter of the Dirichlet process prior. It is well-known (see, e.g. [6]) that the maximum likelihood estimator of ϑ can be obtained by solving, with respect to ϑ , the equation $K_n = \sum_{1 \leq j \leq n-1} \vartheta / (\vartheta + j)$. This method obtained encouraging results but underestimation for small sample size.

3 Illustrations

To compare the performance of the estimators of equations 2, 3, 4 and 5 we rely on some simulations. We fix the population size to $\bar{n} = 10^6$ and consider a sample of $n = 10^5$, representing 10% of the total. Given a fixed size C of the cells at population level, we generate the probability distributions over $(p_j)_{j=1}^C$ according to the Zipf distribution, the uniform distribution and the uniform Dirichlet distribution.

We produce three tables, each according to an increasing choice of the number of cells: $C = 3 \cdot 10^5$ (Table 1), $C = 6 \cdot 10^5$ (Table 2) and $C = 9 \cdot 10^5$ (Table 3). Each column of the tables corresponds to a different choice of the distribution for $(p_j)_{j=1}^C$: the Zipf distribution with parameter $s = 0.2, 0.5, 0.8, 1$, the uniform distribution, the uniform Dirichlet distribution with parameter $\beta = 0.5, 1$. In the first row of each table we have reported the true values of the disclosure index, while the other rows contain the estimates obtained with: i) the naive nonparametric estimator $\hat{\tau}_1^{\mathcal{N}}$; ii) the Bayesian nonparametric estimator $\hat{\tau}_1^{\mathcal{D}}$; iii) the parametric empirical Bayes estimator $\hat{\tau}_1^B$; iv) the parametric empirical Bayes estimator $\hat{\tau}_1^S$. All experiments are averaged

over 100 iterations. The best estimates in each simulated scenarios are shown in bold.

From Tables 1–3, we observe that the performance of the estimators does not follow a clear pattern but depends on the number of cells C along with the distributions over them. In particular, there is not a clear sign highlighting in which scenarios nonparametric estimators outperform parametric estimators, even though it seems that nonparametric estimators might perform better when the size of the contingency table is relative small. This confirms the intuition of [15] for $\hat{\tau}_1^{\mathcal{D}}$ in the case of experimental data.

	Zipf 0.2	Zipf 0.5	Zipf 0.8	Zipf 1	Uniform	Dirichlet 0.5	Dirichlet 1
True τ_1	2868	4651	7537	7313	2579	4151	4608
$\hat{\tau}_1^{\mathcal{N}}$	6413	5613	3810	2157	6511	4232	5111
$\hat{\tau}_1^{\mathcal{D}}$	18461	12471	5421	2459	19303	7498	10741
$\hat{\tau}_1^B$	30554	27271	13848	5358	30847	22820	26351
$\hat{\tau}_1^S$	28702	23621	9670	3043	29187	17665	22306

Table 1 Estimators of τ_1 for several simulated scenarios, when the size of the table is $C = 3 \cdot 10^5$.

	Zipf 0.2	Zipf 0.5	Zipf 0.8	Zipf 1	Uniform	Dirichlet 0.5	Dirichlet 1
True τ_1	16020	16819	16095	11401	15947	9857	12451
$\hat{\tau}_1^{\mathcal{N}}$	7625	6860	4642	2534	7693	5899	6670
$\hat{\tau}_1^{\mathcal{D}}$	32567	21009	7380	2968	33860	14577	20337
$\hat{\tau}_1^B$	33594	31155	17152	6380	33763	28795	31114
$\hat{\tau}_1^S$	34070	29703	13032	3776	34321	25900	29634

Table 2 Estimators of τ_1 for several simulated scenarios, when the size of the table is $C = 6 \cdot 10^5$.

	Zipf 0.2	Zipf 0.5	Zipf 0.8	Zipf 1	Uniform	Dirichlet 0.5	Dirichlet 1
True τ_1	28976	27049	21933	13794	29483	15635	20281
$\hat{\tau}_1^{\mathcal{N}}$	8076	7406	5082	2729	8138	6729	7371
$\hat{\tau}_1^{\mathcal{D}}$	42337	27383	8651	3246	44104	20789	28307
$\hat{\tau}_1^B$	34628	32675	18977	6942	34772	31235	32935
$\hat{\tau}_1^S$	35957	32338	15028	4196	36234	29837	32804

Table 3 Estimators of τ_1 for several simulated scenarios, when the size of the table is $C = 9 \cdot 10^5$.

4 Issues and Open Problems with Nonparametric Approaches

Nonparametric estimation of τ_1 , intended as not imposing any structure on the composition of the population and of the sample, is a very difficult problem. Being a species sampling problem, it presents the typical difficulties encountered in trying to estimate the number of non-observed species (e.g., [7], [5], [11]). [8], seventy years ago, proved that unbiased estimation of the number of unseen classes is impossible when the sample size is smaller than the maximum number of elements found in any class at population level. Another issue typical of nonparametric approaches is the tendency of overestimation and high variance, especially in presence of small samples. All these problems add up in the microdata setting, where data to which we are interested represent a very small fraction of the population. Because of this, strong modelling assumptions on the data structure have been preponderant in the last thirty years.

[16] sum up the state of the art of nonparametric estimation saying that no inference procedure is available that robustly estimates the number of population uniques or sample uniques that are also population uniques without structural assumptions on the distribution of the population and the sample. The estimator $\tau_1^{\mathcal{D}}$ in equation 5 proposed in [15] obtained encouraging results but underestimation for small sample size. As we have seen in Tables 1 and 2, $\tau_1^{\mathcal{D}}$ performed worse in comparison to $\tau_1^{\mathcal{N}}$.

Considering the analogy with species problems, to answer to Skinner’s question we conjecture that a convenient approach to start with would be the empirical nonparametric Bayes, in the sense of [14]. The use of empirical Bayesian methods in species problem dates back to [7] that used it to derive an estimator of the missing mass. They did not formalize the connection, and this link was only highlighted twenty years later in [5]. The use of this approach could lead to a better performing estimator for small samples. Once found such estimator, it would be interesting to find a lower and an upper bound on the normalized mean squared error.

References

1. BETHLEHEM, J.G., KELLER, W.J. AND PANNEKOEK, J. (1990). Disclosure control of microdata. *J. Am. Stat. Assoc.* **85**, 38–45.
2. CAROTA, C., FILIPPONE, M., LEOMBRUNI, R. AND POLETTINI, S. (2015). Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *Ann. Appl. Stat.* **9**, 525–546.
3. CAROTA, C., FILIPPONE, M. AND POLETTINI, S. (2018). Assessing Bayesian nonparametric log-linear models: an application to disclosure risk estimation. *Preprint: arXiv:1801.05244*
4. EFRON, B. AND MORRIS, C (1973). Stein’s estimation rule and its competitors - an empirical Bayes approach. *J. Am. Stat. Assoc.* **68**, 117–130.
5. EFRON, B. AND THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63**, 435–447.

6. FERGUSON, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Ann. Stat.* **1**, 209–230.
7. GOOD, I.J. AND TOULMIN, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43**, 45–63.
8. GOODMAN, L.A. (1949). On the estimation of the Number of Classes in a Population. *Ann. Math. Stat.* **Volume 20, Number 4**, 572–579.
9. MANRIQUE-VALLIER, D. AND REITER, J.P. (2012). Estimating identification disclosure risk using mixed membership models. *J. Am. Stat. Assoc.* **107** 1385–1394.
10. MANRIQUE-VALLIER, D. AND REITER, J.P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *J. Comput. Graph. Stat.* **23** 1061–1079.
11. ORLITSKY, A., SURESH, A.T. AND WU, Y. (2017). Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **113**, 13283–13288.
12. REITER, J.P. (2005). Estimating risks of identification disclosure in microdata. *J. Am. Stat. Assoc.* **100**, 1103–1112.
13. RINOTT, Y. AND SHLOMO, N. (2006). A generalized negative binomial smoothing model for sample disclosure risk estimation. In *Privacy in Statistical Databases. Lecture Notes in Computer Science*, Springer, Berlin.
14. ROBBINS, H. (1956). An empirical Bayes approach to statistics. *Proc. 3rd Berkeley Symp.*, **1**, 157–163.
15. SAMUELS, S.M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *J. Off. Stat.* **14**, 373–383.
16. SKINNER, C.J. AND ELLIOT, M.J. (2002). A measure of disclosure risk for microdata. *J. Royal Stat. Soc. B* **64**, 855–867.
17. SKINNER, C., MARSH, C., OPENSHAW, S. AND WYMER, C. (1994). Disclosure control for census microdata. *J. Off. Stat.* **10**, 31–51.
18. SKINNER, AND SHLOMO, N. (2008). Assessing identification risk in survey microdata using log-linear models. *J. Am. Stat. Assoc.* **103**, 989–1001.

Technologies and data science for a better health both at individual and population level.

Two practical research cases.

Tecnologie e data science per una salute migliore sia a livello individuale che di popolazione.

Stefano Campostrini¹ and Lucia Zanotto¹

Abstract In many fields data are produced at any level and health is not an exception. Every day, each person produces data using personal devices, requiring and accessing to health services or buying drugs. The analysis of these data can drive better public health decision both at individual level and at the population one. Two following research cases tried to provide some answers: ADAPT, “Accessible Data for Accessible Proto-Types in social sector”, funded by Smart Cities, examined how better allocate resources for disabled individuals, also considering home automation solutions; the second one concern a national survey, PASSI, “Progress in the Italian Health Local Units”, which aims to support public health decision making at local, regional and national level.

Abstract

In moltissimi settori, la mole dei dati prodotta è notevole e la salute non è un’eccezione. Ogni giorno, adoperando i dispositivi elettronici, accedendo ai servizi ospedalieri, vengono prodotte informazioni che possono essere analizzate per migliorare i servizi offerti sia a livello individuale che collettivo. Un primo tentativo per raggiungere questi obiettivi sono le seguenti ricerche: la prima, ADAPT, “Accessible Data for Accessible Proto-Types in social sector”, un’indagine finanziata da Smart Cities, è finalizzata al miglioramento della gestione delle risorse per persone disabili anche impiegando le nuove tecnologie in campo di domotica; il secondo studio riguarda l’analisi dei dati dei sondaggi PASSI, “Progress in the Italian Health Local Units”, e mira all’ottimizzazione del processo decisionale sulla salute pubblica a livello locale, regionale e nazionale.

Key words: big data, data science, surveillance systems.

¹

Università Ca’ Foscari Venezia

1 Introduction

Data are all around us. In many fields, data are produced at any level, and health is not an exception. Every day data are generated by using personal devices, accessing to health services, participating in research or simply filling forms for requiring health services, buying drugs at pharmacies and at supermarkets. Analyses of these data can lead to better public health decision at both individual and population level. We would like to discuss few aspects related to the potentialities of these novel approaches, starting from two research cases. The first (individual level), ADAPT (“Accessible Data for Accessible Proto-Types in social sector”) funded by Smart Cities, is a private-academic research which aim is to use smartly data on disabled individuals to better allocate resources for them, supporting services also with home automation solutions. The second study (population level) concerns a national surveillance system, PASSI (Progress in the Italian Health Local Units), which, through monthly surveys, aims to support public health decision at local, regional and national level.

2 ADAPT, “Accessible Data for Accessible Proto-Types in social sector”

The first study we would like to present is a joint private and public project funded by the Italian Ministry of Research, included in a more general program linked to EU projects and related funds: “Smart Cities and Communities and Social innovation”, aimed to increment researches of private companies supported by universities and to develop new instruments and technologies useful to increase citizens’ quality of life. The project, Accessible Data for Accessible Proto-Types in social sector (ADAPT - <http://www.adapt-smartcities.eu/>) presents two main objectives: one related to the analysis of the information system data in order to simplify and enhance the efficacy of the decision making process and the second one regarding the experimentation of home automation technologies (integrated with information systems) to increase the quality of life of disabled persons recipient of public services. The research started in 2013 and, now, the phase of the pilot project has been closed. It has involved two companies specialized in technologies, two private companies providing social and health assistance for elderly and disabled persons, and two universities (University of Venice, Ca’ Foscari, and University of Palermo).

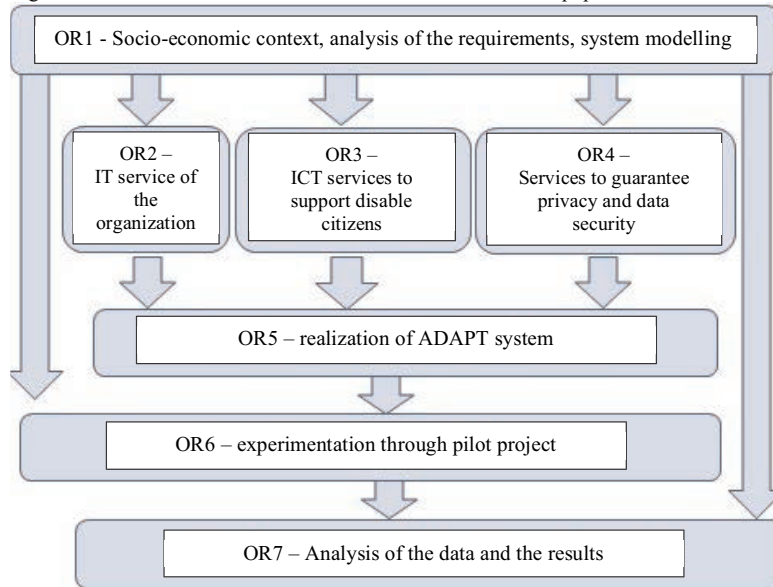


Figure 1: Organization of the project ADAPT

Although not fully implemented (and this is another interesting part of the story), the project has a challenging purpose: trying to obtain as much information as possible from exiting databases to help the decision making process, particularly at individual level. Combining this data with home automatized solutions can help to achieve an integrated system to better satisfy patients' needs. To describe this project we will start from its declared objectives, then we will report very quickly its story, finally we will discuss if the aims has been achieved and how these new solutions can really improve population health.

2.1 *ADAPT objectives*

In order to guarantee that every citizen takes advantages of social, health, economic and cultural actions, policies need to be supported by infrastructures that can:

- guarantee the continuous availability of accurate, timely and relevant information;
- define and manage effective social-welfare paths through the involvement of multi-disciplinary skills and multi-professional profiles;
- ensure the presence of technologically advanced living environments, which can guarantee personalized services set on the real needs of every citizen;
- monitor the spending to better employ the resources.

ADAPT aims to design, implement, test and evaluate technological solutions that support the definition of appropriate and personalized socio-health policies, both with IT support and new perimeter housing technologies, through the creation of a home and territorial RSA prototype.

ADAPT aims to implement and integrate an infrastructure to aid the Electronic Healthcare (unique) File (EHF), extending the EHF with social information and mechanisms to help the management of integrated social-assistance processes, allowing:

- the classification of citizens based on real social assistance needs, optimization of the resources available for their support;
- care continuity, in terms of living environments and social context;
- the modification of living contexts to the improve conditions of partially self-sufficient or non-self-sufficient citizens, in order to guarantee technologically advanced environments dedicated to the inclusion and maintenance of autonomy, without reducing the effectiveness of assistance and rehabilitation.

The final aim is define new welfare models, which allow the achievement of efficiency in services with a more productive use of resources.

2.2 *The ADAPT project development*

Thanks to an agreement signed in 2015, ADAPT project has chosen Veneto as the field for implementation and experimentation in two Local Health Units (LHU-the Region was organized in 20 LHU, reduced to 9 in 2018).

The story of the project implementation has been very troubled, because of bureaucratic problems (on which we will not linger over), the effective capabilities of exiting databases and information system and the possibility to share and match their information. Bureaucracy is always an obstacle to any innovation. Particular concerning data sharing and effective use of information, usually the obstacles come from organizational issues, rather than the development and technical implementation of computer system. ADAPT has not been an exception. Notwithstanding the agreements signed, the realization of the project has been slow and limited.

So the main questions are: do we lack of technologies? Do we lack of data? Do we lack of abilities to analyze data? Or, technology and organizational innovation should go hand in hand?

2.3 *Brief discussion on the ADAPT case*

The increasing needs of the population caused by the demographic effect of ageing leads to the necessity of seeking for smart solutions and decisions. Technology can offer new information and other sources are theoretically available. The (practical) story coming from ADAPT case seem to indicate that we lack of the capability at organizational level mainly, then at data analysis level (because we have not invested enough in analysis capabilities), and finally at information system development: too often, also well-developed organizations, are organized compartmentalized, unable to communicate with one other. ADAPT certainly has produced few interesting solutions on the latter aspects, while for the others it has offered mainly theoretical contributions and shown the key points to change in order to allow an effective implementation of these new ideas and approaches.

3 PASSI, “Progress in the Italian Health Local Units”: a mountain of data for public health decision making

The second study we would like to discuss is, again, an Italian case: the first Behavioral Risk Factor Surveillance (BRFS) national system developed in Europe and globally recognized [8]. The peculiarity of this surveillance system [3] concerns the continuity of the data collection that is more frequent than the usual repeated National Health Surveys, which plans a survey every 3-5 years. This peculiarity produces a unique data stream, analyzable with both the standard statistical tools and more sophistication methods, which can address, as we will see, several complex issues. Moreover, the link with other types of data sources and the employing of “data science” approach could lead to much more information for several complex questions (such as longevity, morbidity and mortality) that could better address public health policies.

3.1 *Aims of the PASSI surveillance system*

The surveillance started from the need to monitor the Italian population health between 18 to 69 years of age. A pilot project was launched in 2007 involving the Local Health Units (LHU) of most of the Italian regions, and in few months has covered more than 90% of the Italian population and data collection is still ongoing (for all the information about methods see [2]). The purpose is collecting longitudinal data systematically to observe citizens’ habits and life style and the impact of new policies in the modification of risky behaviors. The information obtained are mainly useful to those who plan, realize and evaluate public health interventions.

Since the study is realized with the support of each single LHU, which handles the monthly interviews, the representativity of the local areas is ensured. This

characteristic also guarantees prompt reports helpful to analyze the trends of particular phenomena as smoke, alcohol abuse, physical activities, diet and obesity comparing the differences between several territories. The flexibility of the data collection method allows to quickly modify the questionnaire to better answer to the general and specific aims of the survey. Another important aspect of the research is the monitoring of citizens' opinion on their health status and their access to preventive services: the connection between the point of view of the population and health services is an important aspect to evaluate.

3.2 Key challenges, present and future opportunities offered by a data science approach

The potentiality of the data collected by PASSI are relevant because the information collected cover a time window of important transformation of health care, longevity and public health policies. Moreover, the possibility to analyze the differences between areas is interesting to better understand the features of the territories and implement strategies to increase citizens' health levels. The complexity and the size of the records, due to the sizeable number of individuals interviewed and the amount of variables covered, need new statistical approaches. Descriptive analyses, useful to understand the trends and the general tendencies, are only the first step of the research. The interesting district division of the LHU can provide much knowledge about life style and habits per area that can be studied employing spatial statistic techniques. These results can help to evaluate the differences among Italian regions, urban and rural areas, application and efficacy of the national (federal) healthcare system.

Another great opportunity offered by this type of data, which increase every day, is the possibility to connect information regarding morbidity and health status with longevity and causes of death. Since PASSI is a cross-sectional survey, the single individual cannot be followed during his/her life, not offering information regarding ageing or longevity. But other datasets, coming mainly by public (census) registers, can provide information about mortality and causes of death. The problem is how to connect these data sources. An achievable solution could be the probabilistic record linkage: using the values of the observables variables it is possible to merge the individuals of the different datasets into one.

The synthesis of mortality, life styles and socio-economic status can open new prospective to the study of ways for better and longer life. In the last hundred years, the adult distribution of deaths has shown a shift [4] and a compression [5], whose consequences are the increase of the life expectancy at birth [7] and the reduction of the variability around the model age at death. However not all the individuals have been benefit of these changes in the same way: an increase, albeit slight, of the percentage of deaths regarding premature mortality is also observed [9]. Recently an association between cancer and premature mortality has been shown [6], as a stop, or even a reduction of life expectancy in some sub-groups of population. Moreover, the sociocultural status is an important aspect for longevity: globally, persons with

Technologies and data science for a better health both at individual and population level
 higher educational levels have a longer and healthier life. The greater differences are observed around the adult age, where the most common cause of decease is cancer. The systematic study of the connection between causes of death and individual characteristics can answer several questions about the relationship between longevity, risk factors and health policies. Application of more sophisticated statistical tools to these kind of data are promising [1] and a research project with the above mentioned aims has been recently funded by the Italian Ministry of Research (PRIN 20177BRJXS)

4 Final remarks

Health is perhaps the field which could higher benefit form advances in technologies and data science capacities. Progress in medicine and in health systems relies on the implementation of new technologies; data are more and more available and their analyses can help both to find better medical responses, and more effective public health policies and health system organization for the benefit of all the population.

Still, as we have seen in the practical example here quickly discussed, some conditions remain: innovation should be also applied to organizational level, New wine, as known, cannot be put in old wineskins.

Privacy and data security, then, have been raised as reasons against data linkage. Certainly their value cannot be dismissed, but these cannot be an untouchable constrain for research: there are ways to let linkage and analysis respecting privacy and data security.

So, the challenge is open. Much still remains to be done in the health field, and this cannot be left only to private companies that, through data science, study more effective personalized medicine (which not everyone can afford it), while health disparities are increasing everywhere. Innovative technologies and data science can offer better solution for the whole population. The understanding of the causes of different health levels on sub-groups of citizens can lead to a reducing of heath inequalities, an increasing of good health and longevity.

References

1. Assaf S, Campostrini S, Xu F, Crawford CG. Analyzing Behavioral Risk Factor Surveillance Data Using Spatially and Temporally Varying Coefficient Models. *Journals of Royal Statistical Society*, **179**, 1, 153–175 (2016)
2. Baldissera, S., Campostrini, S., Binkin, N., Minardi, V., Minelli, G., Ferrante, G., & Salmaso, S. Peer Reviewed: Features and Initial Assessment of the Italian Behavioral Risk Factor Surveillance System (PASSI), 2007-2008. *Preventing chronic disease*, 8(1) (2011)
3. Campostrini, S., McQueen, D., Taylor, A., & Daly, A. World alliance for risk factor surveillance white paper on surveillance and health promotion. *AIMS public health*, 2(1), 10. (2015)
4. Canudas-Romo, V. The modal age at death and the shifting mortality hypothesis. *Demographic Research*, 19, 1179-1204 (2008)
5. Lan Karen Cheung, S., Robine, J. M. Increase in common longevity and the compression of mortality: The case of Japan. *Population studies*, 61(1), 85-97 (2007)

6. Mazzuco S., Zanutto L., Pastrello C. A Closer Inspection to Rising Premature Mortality in France Using Causes-of-Deaths Data, European Population Conference, Brussels (Belgium) (2018)
7. Oeppen, J. and Vaupel, J. W. Broken limits to life expectancy. *Science*, 296(5570):1029–1031 (2002)
8. WHO, Integrated surveillance of Noncommunicable Diseases (iNCD) Copenhagen (2015).
9. Zanutto L, Canudas-Romo V., Mazzuco S. Evolution of Premature Mortality, European Population Conference, Mainz (Germany) (2016)

Temporal sentiment analysis with distributed lag models

Analisi temporale del “sentiment” con modelli a lag distribuiti

Carrannante M., Mattera R., Misuraca M., Scepi G., Spano M.¹

Abstract Several authors have shown better results in forecasting economic variables by considering the sentiment values in their models. Few studies have focused on the identification of the causes which explain opinions and beliefs. In this paper, we propose a methodological framework based on Distributed Lag (DL) models in order to identify dynamic causal effects in the case of temporal aggregation of sentiment values.

Abstract *In letteratura molti autori hanno introdotto nei modelli di previsione economica i valori del sentiment dimostrando di ottenere così delle previsioni migliori. Pochi sono gli studi che hanno invece come obiettivo l'identificazione delle cause economiche che possano spiegare valori di sentiment. In questo lavoro, si propone un approccio metodologico, basato su modelli a lag distribuiti, con l'obiettivo di identificare gli effetti causali dinamici che si manifestano in presenza di valori del sentiment aggregati in ordine temporale.*

Key words: semantic polarity, social media, causality, dynamic models, textual data

1 Introduction

¹ Maria Carannante, University of Naples “Federico II”. Email: maria.carannante2@unina.it
Raffaele Mattera, University of Naples “Federico II”. Email: raffaele.mattera@unina.it
Michelangelo Misuraca, University of Calabria. Email: michelangelo.misuraca@unical.it
Germana Scepi, University of Naples “Federico II”. Email: germana.scepi@unina.it
Maria Spano, University of Naples “Federico II”. Email: maria.spano@unina.it

Forecasting economic aggregates is an important topic in applied statistics. Many studies were developed in order to obtain the most accurate predictions. The sentiment causality issue appeared only recently in the economic literature. Individual sentiment became an essential additional variable for forecasting economic aggregates, especially after the growing usage of social networks.

In this framework, most of the papers focus on the identification of the effect of sentiment on economic aggregates, whereas very few studies the reverse relationship (see par. 2). In identifying the effects of sentiments on macroeconomic variables, it is necessary to refer to expert opinions and economic agents. In the opposite case, it is interesting to use common people's opinions and sentiments.

In this paper, we propose a causality study in which the dependent variable is the time series of sentiment, while the explicative one is a non-textual sentiment variable related to the economic phenomenon whose sentiment is analysed. Since we expect that the effects of the changes in economic aggregates are not necessarily contemporaneous to the sentiment changes, we propose to use the Distributed Lag (DL) in order to get an identification of the dynamic causal effects.

The paper is structured as follow: after an introduction about the theoretical framework (Section 2), we discuss in detail a methodology which includes both the temporal aggregation of sentiment values and the identification of causal effects, suggesting to use a Distributed Lag model (section 3). The results of case studies are shown in section 4.

2. Theoretical framework

Identifying what drives sentiment is still an unexplored topic since most of the previous papers focused on predictions rather than causality.

A very recent paper related to forecasting economic aggregates using news and sentiment values was developed by Ardia *et al.* (2018), which made their procedure available on R software by a package called "sentometrics". The goal of this paper is to investigate the added value of textual analysis-based sentiment indices for forecasting economic growth, finding that the additional use of news-based sentiment values leads to a significant accuracy gains in forecasting.

However, this paper shows some limitations. In particular, the authors deal with only the variable selection and forecasting, ignoring the possibility of investigating for a causal relationship. Indeed, a critical issue to address is whether economic indices or other exogenous variables influence people's sentiments.

Already in the 1996's paper, Lim & O'Connor (1996) showed the importance of the identification of causal relationships in obtaining better forecasts. Chakrabarty *et al.* (1998) introduced sentiment as an explanatory variable and studied a causal relationship between consumer confidence indexes (used as a proxy of sentiment) and changes in consumer spending. A very similar study was Gelper *et al.* (2007) but from a dynamic perspective. The authors show that the consumer sentiment index can Granger cause (Granger, 1969) future consumption with an average time

Temporal sentiment analysis with distributed lag models

lag of 4-5 months. However also in this study, the sentiment is considered as an explanatory variable and the “sentiment index” is not calculated using data produced by social media.

In a recent paper, Gilitzer & Prasad (2018) investigated the presence of a causal effect of sentiment on consumption also. However, this study focuses on a cross-sectional dimension, without allowing for a dynamic causal effect.

Another study, from a time series perspective, is Zhang *et al.* (2018), which demonstrates that the “daily happiness sentiment”, measured from tweets, Granger causes the market indexes returns. Another similar study is Jiang & Nie (2018), which identifies dynamic causal effects, investigating the relationship between investor sentiment and stock returns for the US economy. However, these papers still identify a causal effect of sentiment on macroeconomic aggregates, ignoring the reverse relationship.

One of the first studies dealing with a similar research question is Graner (1981) where some variables able to explain changes in sentiment values are considered. However, this study has some limitations. First, it is not considered the analysis of textual data in measuring sentiments, using, again, an approach which is not consistent with modern sentiment analysis. Second, the author does not develop a general framework able to quantify in a specific way the presence of a time delayed, dynamic causal effect.

Nevertheless, also more recent papers (e.g. Dehkharghani *et al.*, 2014), that study the effect of some variables in the changes of sentiments, propose only causal rules detection instead of actual identification of causal relationships (for a review of papers see Preethi & Uma, 2015). So it is clear that the causal effect identification issue is still unexplored in literature.

Therefore, in this paper, we propose a strategy which allows us to identify a dynamic regression model in order to capture a dynamic causal effect of an economic time variable on a sentiment time variable.

3. The strategy

Analysing opinions written in natural language is a very interesting research domain, known as sentiment analysis (SA). A large number of papers mention SA in the context of the so-called polarity classification. The main goal is to classify documents written in natural language based on their semantic polarity. This term is commonly used in linguistics to distinguish affirmative and negative forms. The calculation of the positivity/negativity of a document (PN-polarity) entails deciding whether the textual content expresses a positive or negative sentiment. If the document is fractioned into sentences, it is possible to first calculate the polarity of each sentence and then the polarity of the whole document. The polarity score of each sentence depends on the lexicon of polarised terms used, while the polarity of the whole document depends on the polarities of its sentences. The PN-polarity is usually quantified by considering a score of -1 , 0 and $+1$ for the negative, neutral and positive polarity, respectively (Liu, Hu, & Cheng, 2005). Some authors have

proposed different scoring systems by defining the polarity not only in terms of the sign but also taking into account the PN-strength of the sentiment (e.g., Nielsen, 2011). In order to study the presence of dynamic causal relationships between polarity scores and other explanatory economic variables, we need to consider their temporal variations. Most of the economic variables have a time series representation, so it is necessary to transform the polarity scores by aggregating the different values on a temporary basis.

For computing the polarity scores, we follow the approach of Balbi et al. (2018). After pre-processing the texts, the polarity is first calculated at a sentence-level then summarised at a document-level. Considering the date in which the text is published, the polarities are averaged on the same temporal interval and represented in terms of time series. Once the time series of sentiment values are constructed, following the previous papers it is possible to specify the following relationship:

$$y_t = \alpha + \gamma x_t + \beta s_t + \varepsilon_t \quad (1)$$

where y_t is the economic aggregate, x_t a vector of non-textual sentiment variables, s_t the sentiment time series extracted by a temporal aggregation procedure and ε_t the general error term.

We are interested instead in the identification and estimation of a completely different relationship as in the following way:

$$s_t = \alpha + \beta y_t + \varepsilon_t \quad (2)$$

where the sentiment is the dependent variable, y_t is a non-textual sentiment variable relating to the phenomenon whose sentiment is analysed and ε_t is the error term.

The first step is to verify the existence of a causal relationship between (1) and (2). For this purpose, it is possible to perform a causality test as in Granger (1969).

In this way, if we say that y_t “G-causes” s_t , it has sense to develop a framework in which it is possible to estimate the strength of this relationship.

Moreover, since it is not necessarily true that the effect of (1) on (2) is contemporaneous, we need to specify a model which allow us to capture a dynamic causal effect.

The Distributed Lag model is a dynamic model with very useful properties for our purposes, allowing for the presence of an autoregressive structure for the explanatory variable. The Distributed Lag model can be defined, consistently with our notation, as follow (Hendry et al., 1984):

$$s_t = \alpha + \beta(L)y_t + u_t = \alpha + \sum_{i=0}^{n-1} \beta_i y_{t-i} + u_t \quad (3)$$

In (3), the individual coefficients β_i are sometimes called “lag weights” and they explain how y_t affects s_t over time. Indeed, looking at the partial derivatives in this model, the delayed impact of y_t on s_t become clear. If we consider the quantity:

$$\frac{\partial s_t}{\partial y_t} = \beta_0 \quad (4)$$

the (4) represents the contemporaneous or immediate impact of y_t on s_t . Moreover, the derivative respect to the first lag, called β_1 , represents the delayed impact of one lag shifting in the past and so forth. Therefore, the last model can capture a dynamic or delayed impact of y_t on s_t .

The proposed strategy is based on a finite lag model. An incorrect specification of the number of lags to estimate can lead to an incorrect identification. In order to solve this problem, we use an empirical procedure to find the number of lags to include in the model. Firstly, several DL models are estimated. Secondly, the one with the lowest AIC and MASE values is chosen. AIC and MASE are two statistical instruments to select the best model, that relates to different properties of the models. AIC measures the loss of information by using a given model based on a MLE function, so the best model is the one that loses the lowest amount of information, while the MASE measures the accuracy of forecasts, computing the mean absolute scaled error between the forecasting values and the observed values.

Note that a drawback of this approach can be the multicollinearity generated by the autocorrelation structure of the variables y_t . Even if (1) and (2) are stationary time series, y_t could be highly autocorrelated. This statement can lead to unreliable coefficient estimates with large variances and standard errors. Therefore, the identification with the Distributed Lag model is more suitable when the exogenous variable respects these conditions.

4. Case studies

We present some results of two case studies on both economic and financial data. Firstly, we investigate the presence of dynamic causal effects between the growth rate of the United States Industrial Production Index (IPI) and an economic news-based sentiment time series (from 1/1/1995 to 1/12/2014). We collected IPI monthly observations from the Federal Reserve Economic Data website (<https://fred.stlouisfed.org/>), while we extract the sentiment by news published on the Wall Street Journal and The Washington Post.

In particular, we download a collection of news on the American economy from the website <https://www.crowdfunder.com/data-for-everyone> consisting of 4097 documents. We calculate the sentiment values for each document and, successively, we obtain the monthly aggregated values. Figure 1 shows the two time series.

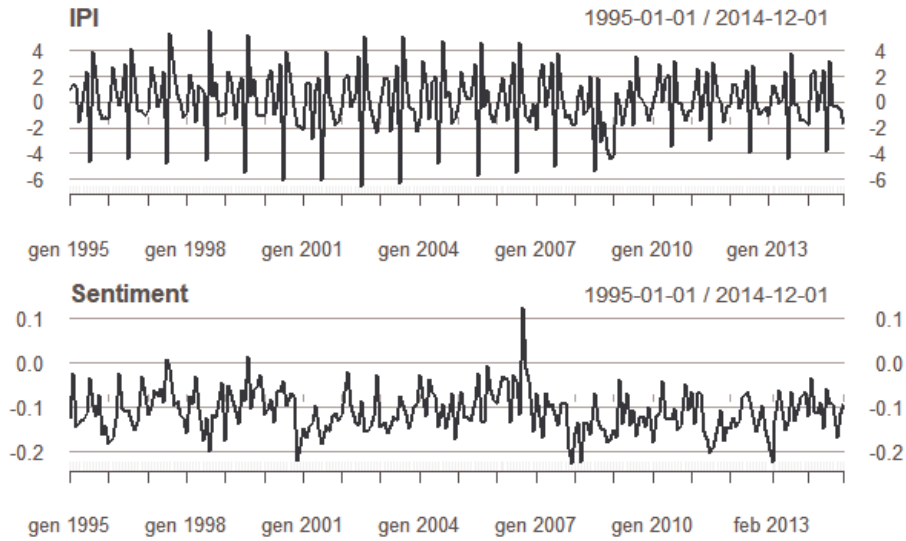


Figure 1: Time series of IPI growth rate and sentiment

Firstly, we must select the number of lags in the model. As we can see in Table 1, information criteria do not lead to an unambiguous decision. For this reason, we performed an iterative procedure and compared the statistical significance of parameters estimation. The best model seems to be the second order one. Therefore we select 2 lags, a congruent solution with AIC criterion.

Table 1: Information criteria

<i>N. of Lags</i>	<i>AIC</i>	<i>BIC</i>	<i>MASE</i>
1	-760.6610	-746.7551	0.82850
2	-763.5940	-746.2327	0.81302
3	-758.5730	-737.7646	0.80726
4	-752.9575	-728.7107	0.80424

We test the Granger causality between the two time series (Table 2). The results show that the sentiment about economic news is caused by the growth rate of the industrial production index. On the other hand, the reverse relationship is statistically significant also.

Table 2: Granger Causality tests

<i>Causality test</i>	<i>N. of lags</i>	<i>F statistic</i>
IPI cause Sentiment	2	4.0425**
Sentiment cause IPI	2	3.3368**

Note: * significance at 10% level; ** significance at 5% level; *** significance at 1% level.

Temporal sentiment analysis with distributed lag models

In Table 3 the estimates of the dynamic causal effects are presented. The results show the presence of a causal delayed effect (only at lag 1 and 2) but not a contemporaneous effect.

Table 3: Parameters estimation

<i>Lags</i>	<i>Estimate</i>	<i>Standard error</i>
Constant	-0.1087873	0.0031428
IPI_t	0.0008596	0.0014336
IPI_{t-1}	0.0043414**	0.0015413
IPI_{t-2}	0.0035803**	0.0014345

Note: * significance at 10% level; ** significance at 5% level; *** significance at 1% level.

Considering the peculiarity of financial time series, we propose another case study. We show here some preliminary results related to a short temporal interval (only 3 years). The financial data are the day values of the spread between Italian BTP and German Bund from 1/1/2016 to 30/12/2018 ([www.https://m.it.investing.com](https://m.it.investing.com)). In the same period, we download a collection of tweets in the Italian language, containing the hashtag “spread”. Figure 2 shows the two time series.

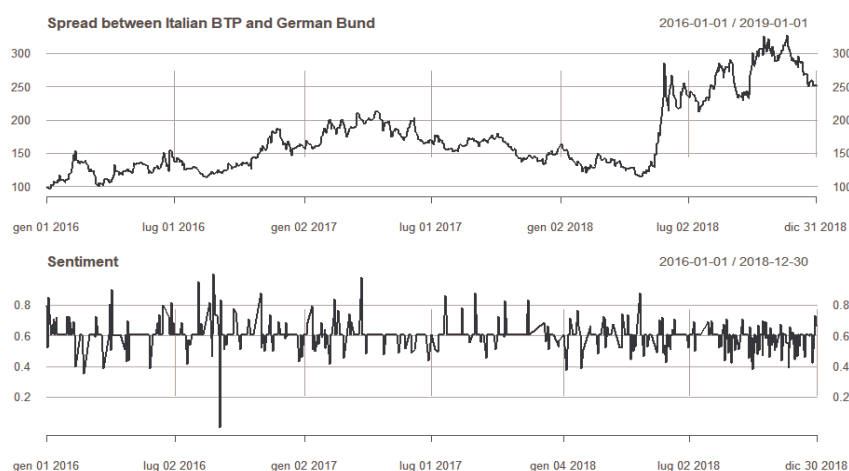


Figure 2: Time series of daily spread and sentiment

We test the Granger causality between the two time series (Table 4). The number of lags is selected following the same procedure used for the first case study. The results show that the sentiment is caused by the spread values. On the other hand, the reverse relationship is not statistically significant.

Table 4: Granger Causality tests

<i>Causality test</i>	<i>N. of lags</i>	<i>F statistic</i>
BTP-Bund spread cause Sentiment	1	4.1308**
Sentiment cause BTP-Bund spread	1	0.1559

Note: * significance at 10% level; ** significance at 5% level; *** significance at 1% level.

The relation is significant, but we do not identify a dynamic causal effect. The reason can be found in the financial time series characteristics. In particular, it is possible to assume that a more sophisticated model is necessary for taking into account the peculiarity of these data. In future developments, we test the model with a bigger data set, and we work on a more appropriate model.

References

1. Ardia, D., Bluteau, K., & Boudt, K.: Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *Int. J. Forecast.* (2018)
2. Balbi, S., Misuraca, M., & Scepi, G.: Combining different evaluation systems on social media for measuring user satisfaction. *Informat. Process. Manag.* 54(4), 674-685. (2018)
3. Chakrabarty, S., Chopin, M., & Darrat, A.: Predicting future buyer behavior with consumers' confidence and sentiment indexes. *Mark. Lett.* 9(4), 349-360. (1998)
4. Dehkharghani, R., Mercan, H., Javeed, A., & Saygin, Y.: Sentimental causal rule discovery from Twitter. *Expert Syst. Appl.* 41(10), 4950-4958. (2014)
5. Gelper, S., Lemmens, A., & Croux, C.: Consumer sentiment and consumer spending: decomposing the Granger causal relationship in the time domain. *Appl. Econ.* 39(1), 1-11. (2007)
6. Gillitzer, C., & Prasad, N.: The effect of consumer sentiment on consumption: cross-sectional evidence from elections. *Am. Econ. J. Macroecon.* 10(4), 234-69. (2018)
7. Granger, C. W. J.: Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* 37 (3), 424-438. (1969)
8. Hendry, D. F. Pagan, A. R., & Sargan, J. D.: Dynamic specification. *Handbook of econometrics*, (2), 1023-1100. (1984)
9. Jiang, Y., Mo, B., & Nie, H.: Does investor sentiment dynamically impact stock returns from different investor horizons? Evidence from the US stock market using a multi-scale method. *Appl. Econ. Lett.* 25(7), 472-476. (2018)
10. Lim, J. S., & O'Connor, M.: Judgmental forecasting with time series and causal information. *Int. J. Forecast.* 12(1), 139-153. (1996)
11. Liu, B., Hu, M., & Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In *Proc. 14th Int. Conf. World Wide Web.* 342-351. (2005)
12. Nielsen, F. Å.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv 1103.2903.* (2011)
13. Preethi, P. G., & Uma, V.: Temporal sentiment analysis and causal rules extraction from tweets for event prediction. *Proced. Comput. Sci.*, 48, 84-89. (2015)
14. Zhang, W., Wang, P., Li, X., & Shen, D.: Twitter's daily happiness sentiment and international stock returns: evidence from linear and nonlinear causality tests. *J. Behav. Exp. Financ.*, 18, 50-53. (2018)

A statistical investigation on the relationships among financial disclosure, sociodemographic variables, financial literacy and retail investors' risk assessment ability

Indagine empirica sulle relazioni tra prospetti per la diffusione di informazioni finanziarie, variabili sociodemografiche, educazione finanziaria e abilità di valutazione del rischio

Rosella Castellano, Marco Mancinelli and Pasquale Sarnacchiaro

Abstract This research focuses on disclosure of information about the uncertain future performance of non-equity financial products. In particular, it looks at two schemes, namely What-if and Probabilistic scenarios. For this purpose, a consumer testing experiment was conducted on a sample of 1,130 potential investors stratified according to ISTAT standard to explore: how different schemes influence investor's risk perception and whether financial literacy and sociodemographic variables are drivers of risk assessment. The analysis was performed using multilevel regression models.

Abstract *Questa ricerca si concentra sui prospetti informativi utilizzati per descrivere le performance aleatorie future di prodotti finanziari di tipo obbligazionario. In particolare, il lavoro considera due schemi informativi*

¹ Rosella Castellano, Dipartimento di Scienze Giuridiche ed Economiche, Università degli studi di Roma Unitelma Sapienza; email: rosella.castellano@unitelmasapienza.it

Marco Mancinelli, Dottorato in Economia e Finanza, Sapienza Università di Roma ; email: marco.mancinelli@uniroma1.it

Pasquale Sarnacchiaro, Dipartimento di Scienze Giuridiche ed Economiche Università degli studi di Roma Unitelma Sapienza; email: pasquale.sarnacchiaro@unitelmasapienza.it

potenzialmente utilizzabili: il metodo What-if e quello degli Scenari Probabilistici. A tal fine, è stato condotto un esperimento su un campione di 1.130 potenziali investitori stratificati in base ai dati ISTAT per esplorare: in che modo i diversi schemi influenzano la percezione del rischio degli investitori e se l'alfabetizzazione finanziaria e le variabili anagrafiche rappresentano fattori trainanti della valutazione del rischio. L'analisi è stata eseguita utilizzando modelli di regressione multilivello.

Key words: Multilevel Regression, Risk Perception, Financial Disclosure, What-If, Probabilistic Scenarios.

Introduction. The empirical breakdown of the rational preference hypothesis in the context of decision-making reveals the limits of classical financial theory. Cognitive theory provides an explanation of some of the biases, and the approach based on these ideas is known as *behavioural finance*. One of the main pillars of behavioural finance are prospect theory and the framing effect (Kahneman and Tversky, 1979, 1981, 1986, 1992), which have been analysed in details in many financial studies (i.e., Bertrand et al., 2010; Hillenbrand and Schmelzer, 2015). The recent financial crisis and the recent cases of securities mis-selling have prompted reflection on the context in which risk assessment and financial decision are made. In this framework, nudging theory has emerged as a tool according to which authorities help people to manage complex situations and avoid legal tricks (Thaler and Sunstein, 2008).

In this context of reference, disclosure is a fundamental issue, since without access to good quality information it is difficult for retail investors to correctly assess the risk of an investment. Investors should be aware of the riskiness of a security and, therefore, it is necessary to provide them with information on uncertain financial cash-flows. Regardless of education, nationality and social condition all potential investors ask the same questions: “What are the risks related to an investment?”. Conscious of the deficiencies in the information delivered to retail investors, both the academy and policy makers have been trying to find the best format to highlight the necessary information for answering this question.

Another of the main factors providing fertile ground for cognitive bias is poor Financial Literacy (FL). Indeed, in-depth world survey (Klapper et al., 2015) shows that people are not usually familiar with very basic financial concepts, with a huge heterogeneity between participants in emerging and advanced economies. As for Italy, only 37% of Italian respondents were financially literate, one of the lowest levels of FL in Europe.

This research focuses on disclosure of information about the uncertain future performance of non-equity financial products. In particular, it looks at two schemes that can be employed to frame the random performances of non-equity financial products: namely What-If (WI) and Probabilistic Scenarios (PS). As they have not been analysed before, we analyze the impact of the two schemes on retail investors' risk assessment ability. For this purpose, a consumer testing experiment was conducted on a sample of 1,130 potential investors stratified according to ISTAT

A statistical investigation between financial literacy and retail investors' risk assessment ability standards in order to explore: how the different schemes influence investor's risk perception and whether FL and sociodemographic variables are drivers of risk assessment. To analyze the relation between information disclosure and risk perception, investors were asked to assess the financial products' riskiness through the two schemes mentioned above. The percentage of respondents that ranked the riskiness of the products correctly was higher when the Probabilistic format was used. Furthermore, frame and FL were the main determinants of risk assessment. This study provides insights into how people actually read and understand financial information, which may prove useful in the design of financial disclosure documents. This research is in line with the approach adopted by some regulators, who are increasingly engaged in the definition of evidence-based rules, and may offer useful insights for the design of more effective representation format. After a brief description of the two methods used to disclose the risk-reward profile of non-equity financial instruments (WI and PS), this paper proceeds as follows. In Sections 3, 4 and 5 the methodology used is discussed, Section 6 presents the results.

2 Possible approaches to disclose risk-reward profile of non-equity financial instruments.

Following discussion about the best format to disclose information, there is a wide consensus that the traditional narrative descriptions of the various factor risks are no longer effective and, therefore, it would be better to use synthetic indicators which are immediately comprehensible to investors and defined in relation to robust and objective quantitative metrics. Despite the progresses made by regulators, the debate about presentation of information continues, particularly in reference to two above mentioned methods: WI and PS. These schemes are applied to two subordinated bonds.

In general, WI analysis is an intensive simulation process whose objective is to capture the behavior of complex systems via specific hypotheses called scenarios (Rizzi, 2016). This method consists, therefore, of calculating the expected return of the product under particular scenarios in order to indicate to a potential investor how the security operates by giving a series of answers to the question "how much could I get if event X happened?". WI scheme is considered the clearest, simplest and least misleading way of representing potential future returns for structured UCITS (see art. 36 of Regulation 583/2010/EU). Guidelines on the WI procedure are set out by the Committee of European Securities Regulators (CESR/10-1318, 2010), although details are not provided. Guidelines require the specification of a set of possible scenarios (namely negative, neutral and positive), which in turn depend on a set of variables. For the two selected subordinated bonds, the main drivers are the default risk and related factors (time to default and recovery rate); the expected return is computed using the internal rate of return. An example of WI table is provided below in Table 1.

The PS scheme is a risk-based method for non-equity products (Minenna et al., 2009; Minenna, 2011), based on possible returns of a financial instrument at maturity. It uses probabilistic tools and is an objective method of determining and

representing the value of synthetic risk indicators that are useful for comparing the various non-equity products available in the global financial market.

In the framework of PS, this paper focuses on the representation of risk-reward profiles. Using the implied risk-neutral probability distribution of financial products and selecting appropriate reference thresholds, one can identify events that are crucial from the perspective of retail investors. The number of events to be disclosed should allow an effective reading of the main statistical features of the distribution (e.g. multimodality and asymmetry) but the maximum number of partitions of the distribution is three. The events should be accompanied by their relative probabilities of occurrence and absolute indicators of performance, namely the mean of the product value in each event (for a complete analysis of this method refer to Minenna et al., 2009; Minenna, 2011). All the data are collected in a table, which provides a synthetic representation in a risk neutral world of the probability distribution of the investment's future value at the end of the recommended investment or holding period. An example of PS table is shown in Table 2.




Table 1: An example of What-if scheme

ILLUSTRATIVE EXAMPLES*	EXPECTED RETURN PER YEAR**
Negative scenario - Issuer defaults before the maturity and only part of the invested amount is paid back.	-3,02%
Neutral scenario – All the coupons are paid regularly but the issuer defaults at maturity.	4,15%
Positive scenario – All coupons and invested capital are paid regularly.	5%

* They are illustrative examples and do not represent a forecast. The shown scenarios may not have an equal probability of occurrence

** If held to maturity

Table 2: An example of PS scheme

EVENTS	PROBABILITY		MEDIUM VALUE IN €*
NEGATIVE performance	66,38%		576
NEUTRAL performance	3,76%		1045
POSITIVE performance	29,86%		1249

* at maturity for an initial investment of 1000 euros

It is worth to highlight here that both Table 1 and 2 represent the analysis under the two schemes of the same subordinated bond issued by an Italian bank in January 2012; it is a seven-year subordinated bond (lower Tier II security) with fixed rate yearly coupons of 5%.

3 Multivariate analysis of risk perception and investment choice.

The main assumption is that the two different risk-reward representations might impact directly on the risk assessment of investment products. In this respect, each participant in the experiment was asked to provide some sociodemographic data, to answer some questions related with basic financial concepts and to assess the risk of the selected two subordinated bonds which were characterized by different levels of risk. Hence, the analysis is focused on the impact of the two frames (described by the variable $T=0$ for WI and $T=1$ for PS) on the binomial variable Risk, Y_i , for the individual i evaluating the risk level of Product I and Product II, labelled as Risk I and Risk II respectively:

$$Y_i = \begin{cases} 1, & \text{Risk I} > \text{Risk II} \\ 0, & \text{otherwise} \end{cases}$$

Although the analysis could be performed by a logistic regression, it is worth moving to a multilevel analysis for purposes of causal inference. In particular, individual- and group-level variations could be taken into consideration in estimating group-level regression coefficients and to model variations among individual-level regression coefficients¹. Therefore, the following research questions should be answered: (i) what is the extent of between-treatment variations in assessing the risk of the two products?; (ii) do individual-level variables such as sociodemographic variables and FL have different effects in different treatments?

4 Model selection

In order to consider the heterogeneity of the individuals' data (level-1) grouped by the two treatments (level-2), the multilevel random effect model is used with the purpose of identifying the factors influencing the risk assessment of the financial products. In particular, Y_{ij} is specified for each i -th individual and the j -th group represented by the treatment as:

$$Y_{ij} = \begin{cases} 1, & \text{Risk I} > \text{Risk II} \\ 0, & \text{otherwise} \end{cases}$$

Since the variable is binomial, the multilevel logistic regression model specification is used. Firstly, the random-intercept specification is identified where a frame-specific (level-2) random intercept is included in the linear predictor thus enabling us to explicitly model the two clusters of the data and their potential unobserved heterogeneity (Goldstein, 2011; Raudenbush and Bryk, 2002). Secondly,

¹ In classical regression, one can do this using indicator variables, but multilevel modeling is convenient when we want to model the variation of these coefficients across groups, make predictions for new groups, or account for group-level variation in the uncertainty for individual-level coefficients.

a random-slope specification is estimated in order to assess whether and to what extent the effect of individuals' characteristics varies across treatments. In this way the assumption of the random-intercept model could be relaxed (i.e. that the slope is fixed across treatments), thus analysing the between-treatment heterogeneity. For the individual i -th in the j -th group, the multilevel odds model is based on the logit link and fitted according to following equations (Hox, 2017; O'Connell *et al.*, 2008; Raudenbush and Bryk, 2002):

$$\text{level 1: } \text{logit}(Y_{ij}) = \text{logit}\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + \sum_{q=1}^Q \beta_q X_{qij} \quad (1)$$

$$\text{level 2: } \beta_{0j} = \gamma_{00} + u_{0j} \quad (2)$$

In these expressions, level-1 equation represents the individual-level model and the level-2 equation defines the treatment-level model. For the i -th individual in the j -th group, Y_{ij} represents the odds for the success category ($y_{ij}=1$) based on $q=1, \dots, Q$ individual-level explanatory variables. The expression on the left-hand side of the equation (1) is the natural log of the odds and it is referred to as the logit distribution. The expression in the middle of equation (1) $\sum_{q=1}^Q \beta_q X_{qij}$ can be interpreted in the same way as the regression coefficients for the multilevel logistic model. The intercept for the j -th group is represented by β_{0j} . The level-2 equation describes how within-group effects may vary according to group-level characteristics. In particular, the intercept is the level-2 variable which is composed by the estimated value (γ_{00}) and the term u_{0j} which is the group-specific random effect that also represents the level-2 error¹.

5 Model specification

Bearing in mind the experimental set, the selection of the variables to be included in the models was carried out by computing association tests according to the nature of each variable (Hosmer and Lemeshow, 2000). Following the common process for developing multilevel analysis (Hox, 2010; Peugh, 2010; Aguinis *et al.* 2013; Sommet and Morselli, 2017), the sociodemographic characteristics of the respondents were introduced in the models as individual-level covariates, namely: gender; age expressed in three classes (20-34; 35-49 and 50-64); area of residence (named as 'Geo') considering four categories² North-West, North-East, Centre and South of Italy; level of education (EDU) in three classes³, marital status and job status. Then, the dimensions related to the investment habits of the respondents such as being updated about financial news (dichotomous variable labelled as 'news') and

¹ $u_{0j} \sim N(0, \sigma_u^2)$

² In line with the categorization made by NIELSEN.

³ The higher degree of education is obtained by three classes: "low" when the respondent got the degree of Licenza Scuola Media or lower; "medium" when obtained at most the degree of Licenza Scuola Superiore; and "high" when the respondent at least gained a university degree.

A statistical investigation between financial literacy and retail investors' risk assessment ability a dummy variable 'exp' to capture their past experience in financial investment, and a categorical measure of FL levels is considered.

After the implementation of an empty two-level model, the analysis continued by building an intermediate model adding explanatory variables, the so-called Constrained Intermediate Model (CIM) so as to assess the variation of the lower-level effects from one cluster to another. Finally, the Augmented Intermediate Model (AIM) is obtained starting from the best CIM: the model is similar to the constrained intermediate model with the exception that it includes the random slope term of FL or for EDUC. Then, the goal is to determine whether the AIM (random intercept-random slope) achieves a better fit to the data than the CIM (random intercept). In other words, our goal is to determine whether considering the cluster-based variation of the effect of the lower-level variable -improves the model. To do so, after gathering or storing the deviance of the CIM and AIM, a likelihood-ratio test is performed, noted LRT. Comparing the random slope model which allows FL to vary across strata and the CIM, the test is statistically significant ($p\text{-value} < 0.05$), hence there is evidence that the effect of FL changes across T.

6 Results

Before considering possible effects of covariates, the multilevel model with only the intercept T is compared to the empty logit model and the LRT statistics has a minimal p-value so that the null hypothesis was rejected¹. Therefore, there is evidence of unobserved heterogeneity at treatment level: as expected, individuals evaluate the risk of Product I and Product II differently under the two frames.

The analysis continued by adding covariates at individual level (level-1) representing respondents' socio-demographic characteristics (such as age, gender and education), their investment habits and FL score, and a parameter related to the experimental setting (R). Some interaction among them are also performed. Since most of the socio-demographic individual characteristics as well as attitudinal to financial investment seem not related to the respondents' propensity to make a correct risk assessment, the only variables resulting statistically significant in the model are the education level and the FL score. The CIM is therefore identified and, after transforming β_{qj} in probabilities, the results are as follows:

Table 3: Please use the *caption* style here

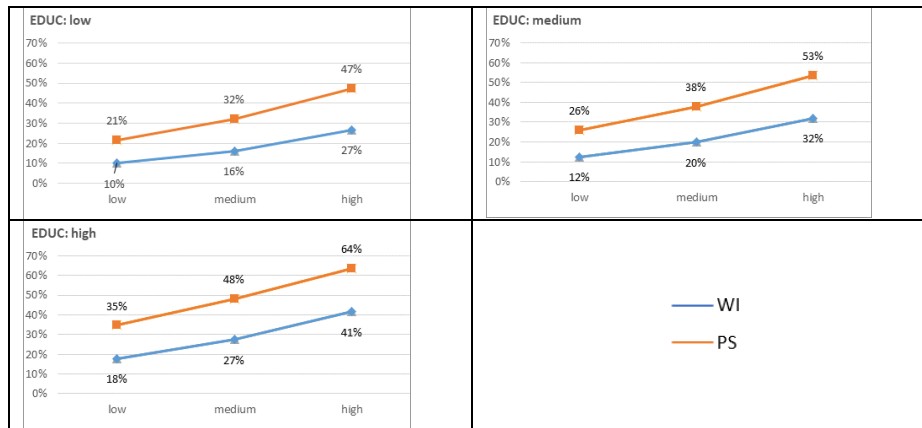
T	Intercept	EDUC medium	EDUC high	FL medium	FL high
0	0.1001193	0.560233	0.6597707	0.6357263	0.766669
1	0.2147556	0.560233	0.6597707	0.6357263	0.766669

The higher are the level of education or the FL score, the higher is the probability of making the correct risk assessment. For example, participants with a high education level assessed riskiness correctly in about 66% more cases than

¹ H0: variance between groups is 0.

individuals with a low education level. This percentage is equal to 77% when comparing individuals with low and high FL. The Figure below reports the predicted probabilities of making the correct risk assessment using the parameters estimates obtained in the CIM. In particular, the results are shown considering the relation between FL and the dependent variable by using EDUC as stratification variable. Firstly, the PS allows potential investor to make the correct risk assessment of the Products more often than the WI do. In particular, participants with low skills (i.e. FL_low, EDUC_low or both of them) seem more advantaged by the PS tables which constantly induce participants to more appropriate answers. Considering the WI treatment, the increase in correct risk assessment is smooth moving from individuals with a low FL score to the ones with high FL score. This is not the case under the Probabilistic representation: a higher level of FL increased significantly the chance of making the correct risk assessment.

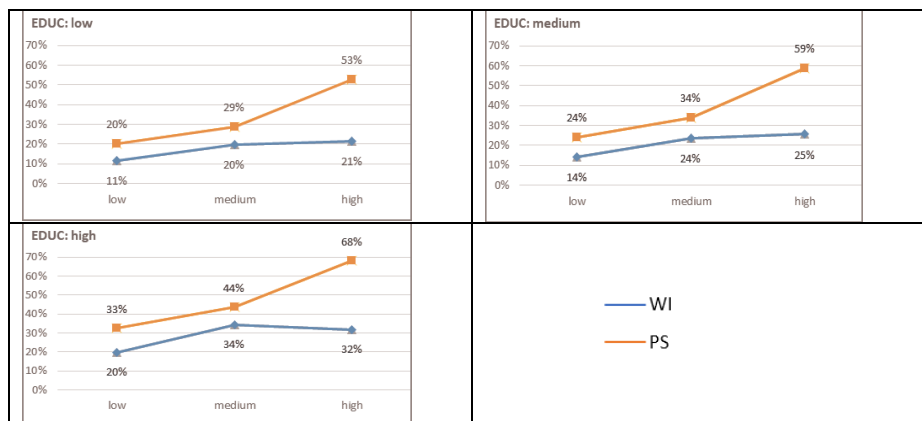
Figure1: CIM predicted probabilities by FL and stratified by education level.



FL is a fundamental factor which acts differently in relation to the two treatments. In general, individuals with higher FL has more chance of making the correct risk assessment. However, respondents with the higher level of FL evaluated Product I and Product II riskiness correctly more frequently under PS scheme than the WI one: individuals with high FL have about 81% more chance than individuals with low FL when PS is used; this relation stays only at 67% under the WI treatment. This phenomenon suggests that: (i) the marginal contribution of higher FL level is strictly positive under the PS; and (ii) since the chance of getting the correct answer is not significantly different between FL_medium and FL_high with respect to FL_low, increasing FL level is not the correct tool to use to potential debias investment decisions when WI representations are used. As for the last point, indeed, even participants with high FL score do not make the correct risk assessment of the financial products. This might highlight the complexity of the WI representation and its unreadable format. All these statements are confirmed by Figure 2 below which

A statistical investigation between financial literacy and retail investors' risk assessment ability shows the marginal contribution of higher level of FL to the chance of achieving the correct risk assessment by using education as stratification variable.

Figure 2: AIM predicted probabilities by FL and stratified by Education level.



The chance of making the correct risk assessment always increase moving from individuals with low FL to the ones with a medium score. The benefit of a further increase of FL score (i.e. moving from medium to high) is not consistent between groups. Indeed, the relationship between FL score and the likelihood of getting the correct risk assessment seems to be always positive under the Probabilistic scheme; whereas under the WI representation this relationship become almost constant at higher level of FL scores.

References

1. Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modelling. *Journal of Management*, Vol. 39(6), pp.1490-1528.
2. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, Vol. 67, Issue 1, pp. 1-48.
3. Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., & Zinman, J. (2010). What's advertising content worth? Evidence from a consumer credit marketing field experiment. *The Quarterly Journal of Economics*, 125(1), pp. 263-306.
4. Goldstein, H., (2011). *Multilevel Statistical Models*, vol. 922. John Wiley & Sons.
5. Hillenbrand, A., & Schmelzer, A. (2015). Beyond information: Disclosure, distracted attention, and investor behaviour. *MPI Collective Goods Preprint*, No. 2015/20.
6. Hosmer Jr., D.W., Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons.
7. Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.
8. Kahneman, D., Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, pp. 263-291.
9. Klapper, L., Lusardi, A., Van Oudheusden, P. (2015). *Financial Literacy Around the World*. Standard & Poor's Ratings Services Global Financial Literacy Survey., Access mode: http://media.mhfi.com/documents/2015-Finlit_paper_17_F3_SINGLES.pdf.

10. Minenna, M. (2011). *A Quantitative Framework to Assess the Risk-Reward Profile of Non-Equity Products*. London, Risk Books.
11. Minenna, M., Boi, G. M., Russo A., Verzella P., Oliva, A. (2009). A quantitative risk-based approach to the transparency on non-equity investment products. *Quaderno di Finanza* n.63, Consob.
12. O'Connell, A. A., Goldstein, J., Rogers, H. J. and Peng, C. J. (2008). Multilevel logistic models for dichotomous and ordinal data. *Multilevel modelling of educational data*, pp.199-242.
13. Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of school psychology*, 48(1), pp. 85-112.
14. Raudenbush, S.W., and Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods (Second Edition)*. Thousand Oaks, CA: Sage Publications.
15. Rizzi, S. (2016). What-if Analysis. *Encyclopedia of Database Systems*, 2nd Edition, L. Liu, T. Özsu (Eds.), Springer.
16. Thaler, R.H., & Sunstein, C.R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
17. Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, New Series*, Vol. 211, No. 4481.
18. Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of business*, S251-S278.
19. Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4), pp. 297-323.
20. Sommet, N., & Morselli, D. (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS. *International Review of Social Psychology*, 30(1), pp.203-2018.

Bayesian Model Comparison based on Wasserstein Distances

Confronto di Modelli Bayesiani tramite Distanze di Wasserstein

Marta Catalano, Antonio Lijoi and Igor Prünster

Abstract Exchangeable processes are extensively used in Bayesian nonparametrics to model exchangeable data. Most common approaches assign a law to the process through the specification of a random measure. When two processes only differ in the law of the random measure, a distance between random measures provides a natural way to compare them. In this work we propose one by relying on the Wasserstein distance. Moreover, we overcome the analytical difficulties of evaluating the distance by developing sharp upper and lower bounds. The specialization of these bounds to Gamma random measures provides the exact value of the Wasserstein distance in terms of the Kolmogorov distance between the base measures. The results are based on a forthcoming work in collaboration with A. Lijoi and I. Prünster.

Abstract *I processi scambiabili sono usati di frequente per modellare dati scambiabili. Nella maggior parte dei casi la legge del processo richiede la specificazione di una misura aleatoria. Quando due sequenze scambiabili differiscono solamente nella distribuzione delle misure, la valutazione di una distanza tra misure aleatorie fornisce un modo naturale di metterle a confronto. In questo lavoro ne proponiamo una basata sulla distanza di Wasserstein. Inoltre, superiamo le difficoltà analitiche tramite la derivazione di limiti superiori e inferiori. La specializzazione dei limiti alle misure aleatorie Gamma fornisce il valore esatto della distanza di Wasserstein in termini di distanza di Kolmogorov tra misure base. I contenuti si basano su un lavoro di prossima pubblicazione in collaborazione con A. Lijoi e I. Prünster.*

Key words: Bayesian nonparametrics, Completely random measures, Increasing additive processes, Wasserstein distance.

Marta Catalano
Bocconi University, Milano, Italy. e-mail: marta.catalano@unibocconi.it

Antonio Lijoi
Bocconi University and Bocconi Institute of Data Science and Analytics (BIDSA), Milano, Italy.
e-mail: antonio.lijoi@unibocconi.it

Igor Prünster
Bocconi University and BIDSA, Milano, Italy. e-mail: igor.pruenster@unibocconi.it

1 Introduction

Consider a generic parametric class of models $\mathcal{M} = \{M_\theta \mid \theta \in \Theta \subset \mathbb{R}^k\}$ assumed to describe or approximate the distribution of n observations (x_1, \dots, x_n) . Many Bayesian inferential procedures rely on a notion of *discrepancy* between models, which is often translated into a distance between random variables. For example, sensitivity to the prior is assessed through a comparison between the posterior distributions, which often amounts to the evaluation of a (pseudo-)distance; see [2] for a review. Moreover, these are also used in model selection [10], variable selection [7], and in general in Bayesian testing when the hypotheses are nested. In this context many authors think that Bayes factors, corresponding to 0–1 losses, may be too restrictive and prefer considering distance-based losses instead [3, 16, 17].

A consistent portion of the Bayesian literature now focuses on the specification and properties of the so-called nonparametric priors, whose large support guarantees reliable predictions and estimations. The analytical tractability of the Dirichlet process [9] opened the way to the study of many nonparametric structures relying on random measures. For example, random measures are often normalized so to provide de Finetti measures [14, 15] or to define priors for density functions [1], which are both used to specify the law of an exchangeable process. In survival analysis, moreover, they provide effective ways to specify the law of random hazard functions [8] or cumulative hazards [6, 11]. In all these cases, the specification of the law for a stochastic process of interest requires the distribution of a random measure. Typical inferential procedures analyse how this distribution is affected by the observed data. Nonparametric analogues of the previous procedures, such as sensitivity assessment and hypothesis testing, could then be based on distances between random measures. Interestingly, to the best of our knowledge there has not been any attempt to define such distances in the Bayesian nonparametric literature. We here propose a way to fill in this gap by exploiting the Wasserstein distance. While simulations of the Wasserstein distance are easily achieved [19], analytical evaluations are generally difficult. This raises the need for analytically tractable and sharp bounds. We achieve such bounds for a wide subclass of random measures, the so-called completely random measures.

The outline of the work is the following. After a brief recapitulation of basic notions about completely random measures and the Wasserstein distance, in Section 3 we provide general upper and lower bounds for the Wasserstein distance between completely random measures. These are expressed in terms of the underlying Lévy measures and are then specialized to Gamma completely random measures in Section 4.

2 Preliminaries

This section will be devoted to the definition of a distance between random measures. We first recall some useful properties of the Wasserstein distance and of completely random measures.

Let \mathbb{X} be a Polish space with respect to a metric d , endowed with the Borel σ -algebra $\mathcal{B}(\mathbb{X})$, and let X_1 and X_2 be \mathbb{X} -valued random elements. The Wasserstein distance of order $p \in [1, +\infty)$ between X_1 and X_2 is defined as

$$W_{p,d}(X_1, X_2) = \inf_{(Z_1, Z_2) \in C(X_1, X_2)} \left\{ \mathbb{E}(d(Z_1, Z_2)^p)^{\frac{1}{p}} \right\},$$

where $C(X_1, X_2)$ indicates the Fréchet class of X_1 and X_2 , i.e. the set of distributions on the product space \mathbb{X}^2 whose marginal distributions on \mathbb{X} coincide with the laws of X_1 and X_2 . In the rest of the paper we will focus on the case $p = 1$ and $(\mathbb{X}, d) = (\mathbb{R}, |\cdot|)$, i.e. the real line with Euclidean distance, and we will denote such distance W . It can be shown that

$$|\mathbb{E}(X) - \mathbb{E}(Y)| \leq W(X, Y) \leq \mathbb{E}(|X|) + \mathbb{E}(|Y|). \quad (1)$$

In particular, the Wasserstein distance is finite when the random variables have finite mean.

Consider the space $\mathbb{M}(\mathbb{R})$ of boundedly finite measures on \mathbb{R} endowed with the weak[#] topology [5], and denote by $\mathcal{M}(\mathbb{R})$ the corresponding Borel σ -algebra. A random measure is a random element on the Borel space $(\mathbb{M}(\mathbb{R}), \mathcal{M}(\mathbb{R}))$. We identify each random measure $\tilde{\mu}$ with its corresponding cumulative process $\{\tilde{\mu}((-\infty, x])\}_{x \in \mathbb{R}}$ and propose the following distance between random measures.

Definition 1. Given two random measures $\tilde{\mu}_1$ and $\tilde{\mu}_2$ on \mathbb{R} we define

$$d_W(\tilde{\mu}_1, \tilde{\mu}_2) = \sup_{x \in \mathbb{R}} W(\tilde{\mu}_1((-\infty, x]), \tilde{\mu}_2((-\infty, x])),$$

where W is the Wasserstein distance. It is easily shown that d_W is a distance on the laws of random measures and we refer to it as the Wasserstein distance between random measures. The distance d_W is finite whenever $\mathbb{E}(\tilde{\mu}_i(\mathbb{R})) < +\infty$, for $i = 1, 2$.

The rest of the work concerns the evaluation of this distance on the so-called completely random measures. A random element $\tilde{\mu}$ taking values in $(\mathbb{M}(\mathbb{R}), \mathcal{M}(\mathbb{R}))$ is a completely random measure (CRM) if, given a finite collection of pairwise disjoint bounded sets $\{A_1, \dots, A_n\}$ in $\mathcal{B}(\mathbb{R})$, the random variables $\{\tilde{\mu}(A_1), \dots, \tilde{\mu}(A_n)\}$ are mutually independent [12]. Every CRM $\tilde{\mu}$ can be decomposed as the sum of three independent components, $\tilde{\mu} \stackrel{d}{=} \mu + \tilde{\mu}_f + \tilde{\mu}_c$, where μ is a deterministic measure, $\tilde{\mu}_f$ is a random measure with fixed atoms and $\tilde{\mu}_c$ is a random measure without fixed atoms. Let $\mathbb{R}^+ = (0, +\infty)$. In particular, for every CRM without fixed atoms there exists a diffuse boundedly finite measure ν on $\mathbb{R}^+ \times \mathbb{R}$ such that

$$\tilde{\mu}_c(dy) \stackrel{d}{=} \int_0^{+\infty} s \mathcal{N}(ds, dy),$$

where \mathcal{N} is a Poisson random measure with intensity \mathbf{v} . This means that \mathcal{N} is a CRM on $\mathbb{R}^+ \times \mathbb{R}$ and, for any $B \in \mathcal{B}(\mathbb{R}^+) \otimes \mathcal{B}(\mathbb{R})$ such that $\mathbf{v}(B) < \infty$, $\mathcal{N}(B)$ is a Poisson random variable with mean $\mathbf{v}(B)$. The corresponding cumulative process $\{\tilde{\mu}_c((-\infty, x])\}_{x \in \mathbb{R}}$ is an increasing additive process on \mathbb{R} with Lévy measures $\mathbf{v}_x(ds) = \int_{(-\infty, x]} \mathbf{v}(ds, dy)$ satisfying

$$\int_0^1 s \wedge 1 \mathbf{v}_x(ds) < +\infty \quad \forall x \in \mathbb{R}.$$

In applications to Bayesian frameworks one is usually interested in CRMs that are infinitely active, i.e. such that $\mathbf{v}_x((0, 1]) = +\infty \forall x \in \mathbb{R}$. Moreover, we point out that $\tilde{\mu}_c$ has finite mean if and only if $\sup_x \int_0^{+\infty} s \mathbf{v}_x(ds) < +\infty$.

3 Wasserstein Bounds for Completely Random Measures

The Wasserstein distance can be easily simulated [19] but it is generally difficult to evaluate analytically. Nonetheless this is an important task since it can be used to quantify, for example, the sensitivity of the model to the prior specification of CRM. Our purpose for this section is to provide a general framework to derive upper and lower bounds for the Wasserstein distance between completely random measures in terms of their corresponding Lévy measure.

We focus on CRMs without fixed atoms, though our results can be extended in a natural way to CRMs with atoms. If $\tilde{\mu}$ is infinitely active, its Lévy measure \mathbf{v} is diffuse and not finite. Thus, for $i = 1, 2$ and $r > 0$ there exists $\varepsilon_{i,r} > 0$ such that

$$\mathbf{v}_{i,x}([\varepsilon_{i,r}, +\infty)) = r.$$

We further define the probability measure $\rho_{i,r,x}$ to be proportional to the restriction of $\mathbf{v}_{i,x}$ on the interval $[\varepsilon_{i,r}, +\infty)$, i.e.

$$\rho_{i,r,x}(ds) = \frac{\mathbf{v}_{i,x}(ds)}{r} \mathbb{1}_{[\varepsilon_{i,r}, +\infty)}(s),$$

which can be shown to be the distribution of the jumps of the compound poisson approximation of $\mu_i((-\infty, x])$; see [18].

Theorem 1. *Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ be infinitely active CRMs with finite mean. Then for every $x \in \mathbb{R}$*

$$\left| \int_0^{+\infty} s (\mathbf{v}_{1,x} - \mathbf{v}_{2,x})(ds) \right| \leq W(\tilde{\mu}_1((-\infty, x]), \tilde{\mu}_2((-\infty, x])) \leq \lim_{r \rightarrow +\infty} r W(\rho_{1,r,x}, \rho_{2,r,x}).$$

Moreover, the limit on the right hand side is finite.

The lower bound is an immediate consequence of (1). As for the right hand side, it can be seen as a generalization of [13], where the authors provide upper bounds on the Wasserstein distance between Lévy processes; see [4] for a proof.

4 Gamma Completely Random Measures

In this section we apply Theorem 1 to evaluate the exact expression of the Wasserstein distance between Gamma completely random measures. We recall that a Gamma CRM $\tilde{\mu}$ of parameter $b > 0$ and base measure α has Lévy intensity

$$\nu(ds, dy) = \frac{e^{-sb}}{s} \mathbb{1}_{(0, +\infty)}(s) ds \alpha(dy).$$

We use the notation $\tilde{\mu} \sim \text{Ga}(b, \alpha)$. The random measure $\tilde{\mu}$ is easily shown to be infinitely active and, if α is a finite measure on \mathbb{R} , it has a finite mean. Moreover, we set $x \mapsto A(x) = \alpha((-\infty, x])$.

Theorem 2. *Let $\tilde{\mu}_i \sim \text{Ga}(b_i, \alpha_i)$, where $0 < b_1 < b_2$ and α_i is a finite measure on \mathbb{R} for $i = 1, 2$. Then,*

1. *If $\alpha_1 = \alpha_2 = \alpha$,*

$$W(\tilde{\mu}_1((-\infty, x]), \tilde{\mu}_2((-\infty, x])) = A(x) \left| \frac{1}{b_1} - \frac{1}{b_2} \right|;$$

2. *If $b_1 = b_2 = b$,*

$$W(\tilde{\mu}_1((-\infty, x]), \tilde{\mu}_2((-\infty, x])) = \frac{1}{b} |A_1(x) - A_2(x)|.$$

Remark 1. Theorem 2 clarifies the sharpness of the bounds derived in Theorem 1. Indeed, in this case the upper and lower bound coincide and can thus be used to derive the exact value of the Wasserstein distance.

Remark 2. Theorem 2 provides an immediate evaluation of the distance between completely random measures defined in Definition 1. By taking the supremum over $x \in \mathbb{R}$, one derives

1. *If $\alpha_1 = \alpha_2 = \alpha$,*

$$d_W(\tilde{\mu}_1, \tilde{\mu}_2) = \alpha(\mathbb{R}) \left| \frac{1}{b_1} - \frac{1}{b_2} \right|;$$

2. *If $b_1 = b_2 = b$,*

$$d_W(\tilde{\mu}_1, \tilde{\mu}_2) = \frac{1}{b} K(\alpha_1, \alpha_2),$$

where $K(\alpha_1, \alpha_2)$ indicates the Kolmogorov distance between two finite measures, namely

$$K(\alpha_1, \alpha_2) = \sup_x |A_1(x) - A_2(x)|.$$

Intuitively, one expects that CRMs with similar parameters are close to each other. Our results confirm the intuition and allow for a precise quantification of the *closeness* in terms of Wasserstein distance.

References

1. Barrios, E. and Lijoi, A., Nieto-Barajas, L. E. and Prünster, I.: Modeling with normalized random measure mixture models. *Stat. Sci.* **28**, 313–334 (2013)
2. Berger, J.O.: An overview of robust Bayesian analysis (with discussion). *Test.* **3**, 5–124. Springer (1994)
3. Bernardo, J. M., Rueda, R.: Bayesian hypothesis testing: A reference approach. *Int. Stat. Rev.* **70**, 351–372 (2002)
4. Catalano, M., Lijoi, A., Prünster, I.: Wasserstein convergence rates of an approximate sampler for time-to-event data. In preparation.
5. Daley, D.J. and Vere-Jones, D.: An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods. Probability and Its Applications. Springer (2002)
6. Doksum, K.: Tailfree and Neutral Random Probabilities and Their Posterior Distributions. *Ann. Probab.* **2**, 183–201 (1974)
7. Dupuis, J. A., Robert, C. P.: Variable selection in qualitative models via an entropic explanatory power. *J. Stat. Plan. Inference.* **111**, 77–94 (2003)
8. Dykstra, R. L., Laud, P.: A Bayesian Nonparametric Approach to Reliability. *Ann. Statist.* **9**, 356–367 (1981)
9. Ferguson, T. S.: A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230 (1973)
10. Goutis, C., Robert, C.: Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. *Biometrika.* **85**, 29–37 (1998)
11. Hjort, N. L.: Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259–1294 (1990)
12. Kingman, J. F. C.: Completely random measures. *Pacific J. Math.* **21**, 59–78 (1967)
13. Mariucci, E. and Reiß, M.: Wasserstein and total variation distance between marginals of Lévy processes. *Electron. J. Statist.* **12**, 2482–2514 (2018)
14. Pitman, J. and Yor, M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900 (1997)
15. Regazzini, E. and Lijoi, A. and Prünster, I.: Distributional results for means of normalized random measures with independent increments. *Ann. Statist.* **31**, 560–585 (2003)
16. Rousseau, J.: Approximating interval hypothesis: p-values and Bayes factors. *Bayesian Statistics 8, Oxford Sci. Publ.* **8**, 417–452 (2007)
17. Salomond, J.-B.: Testing Un-Separated Hypotheses by Estimating a Distance. *Bayesian Anal.* **13**, 461–484 (2018)
18. Sato, K.: Lévy Processes and Infinitely Divisible Distributions. Cambridge University Press (1999)
19. Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., Lanckriet, G. R. G.: On the empirical estimation of integral probability metrics. *Electron. J. Statist.* **6**, 1550–1599 (2012)

Hierarchical Clustering and Dimensionality Reduction for Big Data

Clustering e Riduzione Dimensionale Gerarchici per Dati di Grandi Dimensioni

Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria

Abstract The development of new technologies and methods of data collection produces the necessity to summarise the large quantity of information that is available. Usually, we face a data matrix \mathbf{X} of size $(n \times J)$, corresponding to n statistical units and J quantitative variables, where n and J are very large. Clustering is the analysis which identifies homogeneous clusters of units, thus it might be meant as a way to reduce their dimension. Dimensionality reduction techniques are methods to obtain latent dimensions (less than manifest variables), so they reduce the dimensionality of the variables space. In this paper, we apply *Double Hierarchical Parsimonious Means Clustering* [2] in order to get a simultaneous hierarchical parsimonious clustering of units - aggregated around centroids - and dimensionality reduction of variables - aggregated around components - on *Asia-Europe Meeting* (ASEM) data set. The model is estimated by using the LS method and an efficient coordinate descent algorithm is given. The goodness of fit of the double hierarchical parsimonious trees can be computed to assess the quality of the two hierarchical partitions.

Key words: clustering, dimensionality reduction, big data, hierarchy.

Carlo Cavicchia
University of Rome La Sapienza, Piazzale Aldo Moro 5, Rome,
e-mail: carlo.cavicchia@uniroma1.it

Maurizio Vichi
University of Rome La Sapienza, Piazzale Aldo Moro 5, Rome,
e-mail: maurizio.vichi@uniroma1.it

Giorgia Zaccaria
University of Rome La Sapienza, Piazzale Aldo Moro 5, Rome,
e-mail: giorgia.zaccaria@uniroma1.it

1 Introduction

In recent years, with the data revolution and the use of new technologies, phenomena are frequently described by a huge quantity of information useful for making strategical decisions. A priority for policymakers is having simple statistical methods useful to synthesise all the available information. Different levels of synthesis are required by stakeholders in order to describe properly different phenomena.

Cluster analysis is a field of study which tries to identify homogeneous clusters of units. Hierarchical clustering methods are well-known and widely used for producing a hierarchy of statistical units, clustered in $(n - 1)$ nested partitions. Most of these methods take into account an objects-by-objects dissimilarity matrix, by setting a priori the kind of metrics and linkage in order to measure and update the distance between items, respectively. These clustering methods are heuristic and they do not underpin a model for the dissimilarity data or an objective function that can be optimised. [7] proposed a method that extends *K-means* to the case of hierarchical clustering estimating the objective function via least squares.

Dimensionality reduction methods (e.g. Principal Component Analysis (PCA) and Factor Analysis (FA)) are usually implemented to obtain a straightforward interpretation of the data. These methodologies are sometimes not able to get the real structure of the data and their relationships, i.e. a hierarchical correlation structure.

[3] proposed a hierarchical extension of *Disjoint Principal Component* [4] in order to build composite indicators.

In this paper, we apply the *Double Hierarchical Parsimonious Means Clustering* [2] in order to get a simultaneous hierarchical partitions of units - represented by centroids - and of variables - represented by components - on the *Asia-Europe Meeting* (ASEM) dataset. The aim of this research is to build a composite indicator for ASEM taking into account a hierarchical set of nested partitions of countries.

The paper is organised as follows. In Section 2 the model is presented and in Section 3 it is applied on the ASEM dataset. Finally, in Section 4 some conclusions end the paper.

2 Methodology

In the era of big data, the need to synthesise information is even more crucial. Clustering and dimensionality reduction are considered in order to synthesise large quantity of data. Both for units and for variables, it is worthy to identify clusters or classes of objects that represent homogeneous features. On one hand, the huge amount of data holds much more information than previously and millions of statistical units are available; on the other hand, it becomes necessary to understand if this information might be transformed into statistical knowledge.

The syntheses of objects and variables are usually achieved according to sequential or simultaneous approaches, as the *tandem analysis* or the *Clustering and Disjoint Principal Component Analysis* (CDPCA) proposed by [8], respectively. Many

authors have criticised the former method since it brings about a masking of the taxonomic information of the data. However, the simultaneous approach does not allow to inspect the hierarchical relationships between dimensions of a multidimensional phenomenon, whenever they exist.

In the specialised literature, many methodologies have been developed to simplify the complete hierarchies ([5]) and to build parsimonious trees ([6]), both for units and variables. In case of big data, the parsimony property is fundamental to interpret the results.

[2] studied a new hierarchical simultaneous model-based approach to cluster objects and to identify new latent concepts, each one associated to a group of variables. The methodology is based upon the CDPCA, starting from a fixed number of clusters K and components Q and reducing these values by one at each hierarchical level. Formally,

$$\mathbf{X} = \mathbf{U}_k \mathbf{M}_{kq} \mathbf{V}_q' \mathbf{B}_q + \mathbf{E}_k \quad \forall k = K, \dots, 1, q = Q, \dots, M, \quad (1)$$

where \mathbf{X} is a $(n \times J)$ data matrix - with n statistical units and J quantitative variables-, \mathbf{U}_k and \mathbf{V}_q are the membership matrices for units and variables, respectively, \mathbf{B}_q is the matrix of weights and \mathbf{M}_{kq} is the centroids matrix in the reduced space.

The model (1) is subject to the classical constraints on membership matrices for partitioning and, according to [8], on the reparametrization of the loading matrix \mathbf{A}_q into the product of two matrices, i.e. \mathbf{B}_q and \mathbf{V}_q . Furthermore, a constrain on nested partitions has been added to the model (1).

Eq.(1) represents the reflective part of the model with $Q - M + 1$ hierarchical levels. M identifies the number of the bottom-up level of the hierarchy at which the model becomes formative, i.e. the M components are merged into a unique measure of synthesis, and it is selected according to a statistical test.

The model is estimated in a least-squares semi-parametric framework in which a quadratic loss function is minimised and it is implemented with a coordinate descent algorithm. The latter is efficient in real applications.

3 Application: ASEM Index - International Sustainable Connectivity

Asia-Europe Meeting (ASEM) Sustainable Connectivity Index is aimed at measuring connectivity among countries, people and societies in an economic sense (e.g. transport links, energy, trade,...) and in a social sense (e.g. migration, linkage, cultural connection,...). The data comprises 51 countries¹ - 30 European and 21 Asian - and 49 indicators.

¹ Source: [1].



The indicators are grouped in two indexes, *Connectivity* and *Sustainability*, with 5 and 3 dimensions respectively, as shown in Figure 1.

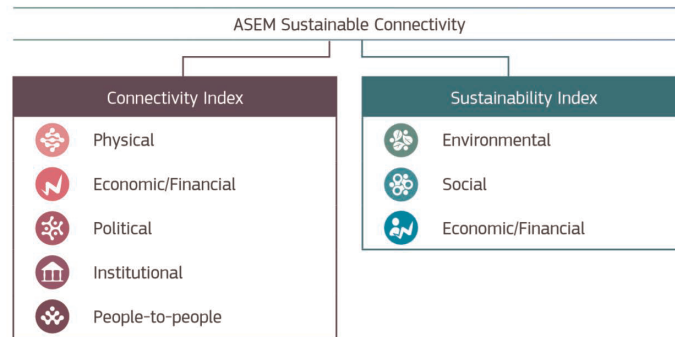


Fig. 1 ASEM Sustainable Connectivity Conceptual Framework¹.

The methodology described in Section 2 has been implemented on this data set for clustering countries and build a composite indicator, i.e. a measure of synthesis, from the 49 indicators. With respect to the construction of the variables hierarchy, two research approaches are defined: confirmatory and exploratory. In the former, the 8 dimensions are fixed (Figure 1), i.e. the partition of the manifest variables at the eighth hierarchical level is constrained. In the latter, all the constraints are relaxed and the initial parsimonious number of variable groups is pinpointed according to the unidimensionality of the components. In both cases, the optimal solution of the model corresponds to 6 clusters of statistical units. Before analysed the results and in order to measure the internal reliability of the two proposed indices, the Cronbach's α has been computed: the Connectivity index has $\alpha = 0.94$ and the Sustainability one has $\alpha = 0.37$. Thus, the former seems to be very consistent, whereas the latter turns out to be not reliable.

In the confirmatory approach, the model (1) underpins the theoretical double composite indicators approach with $M = 2$, whose corresponding components are

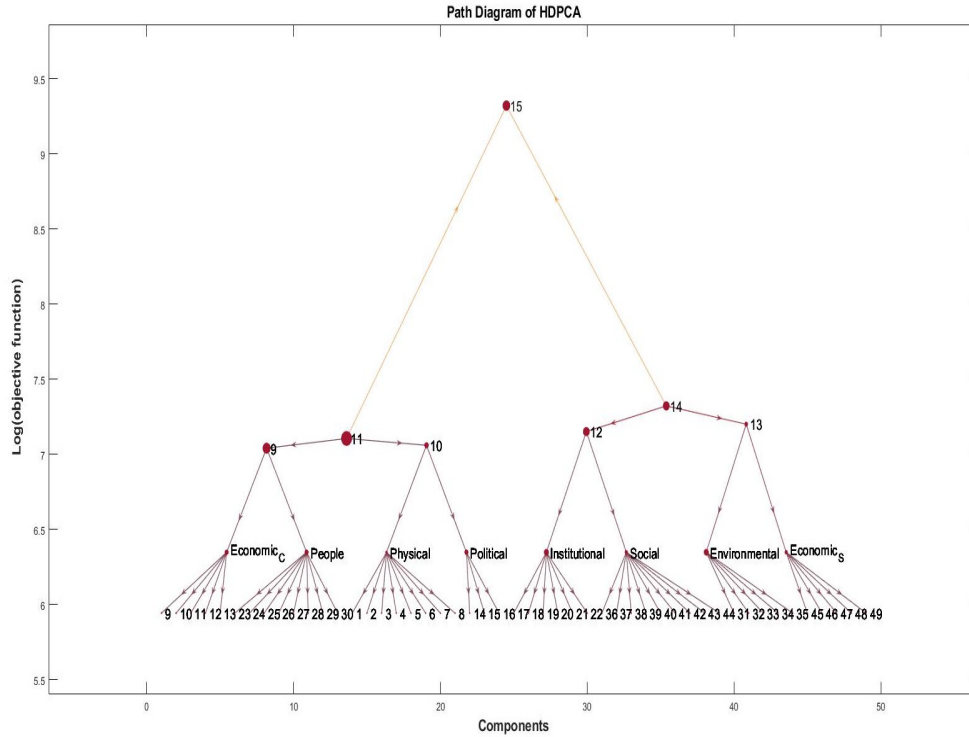


Fig. 2 Conrmatory Approach on ASEM Data Set. Variables Hierarchy.

merged together in a formative way. The partition obtained at this hierarchical level is equal to the one proposed, except for the dimension *Institutional* - which belongs to the Sustainability index group according to the model (1) - as shown in Figure 2. The Cronbach's α shows improvements for this group, passing from 0.37 to 0.67. Moreover, only one dimension is not unidimensional (*Political*) - the unidimensionality is assessed according to the magnitude of the "restricted"² covariance matrix second eigenvalue.

In the exploratory approach, the model (1) pinpoints $Q = 3$ unidimensional components and it underpins again the theoretical model identifying $M = 2$. The three groups are composed by the following variables of the theoretical domains:

- 4/8 *Physical*, 5/5 *Economic/Financial* (Connectivity), 1/3 *Political*, 7/8 *People-to-people*.
- 4/8 *Physical*, 2/3 *Political*, 6/6 *Institutional*, 1/8 *People-to-people*, 1/5 *Environmental*, 9/9 *Social*.
- 4/5 *Environmental*, 5/5 *Economic/Financial* (Sustainability).

² It refers to the manifest variables of the data matrix associated to a component.

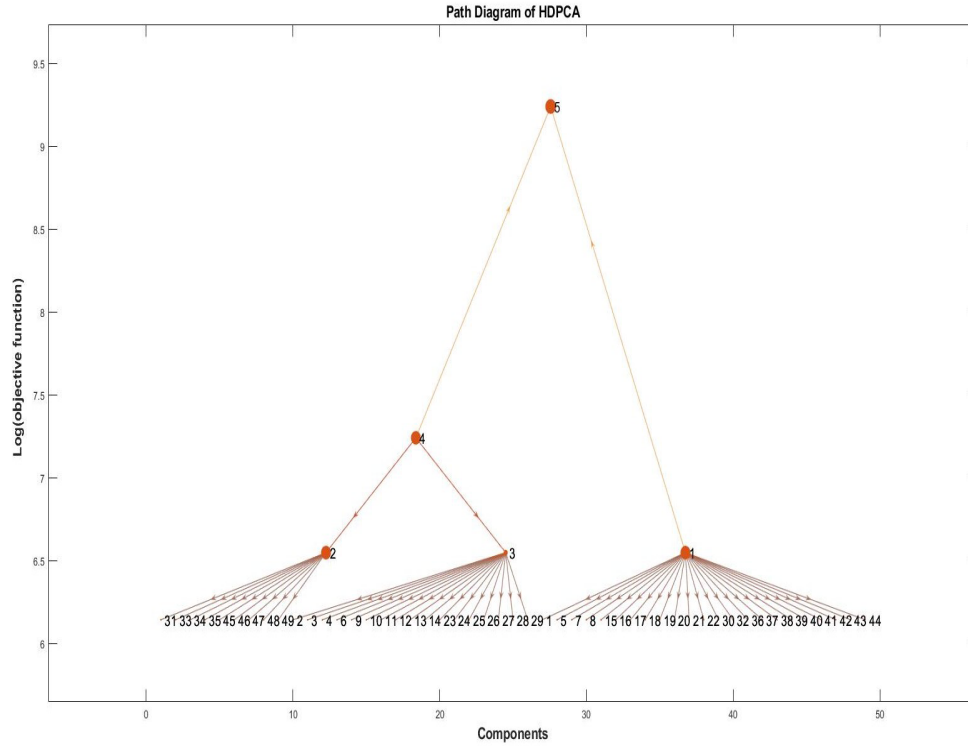


Fig. 3 Exploratory Approach on ASEM Data Set. Variables Hierarchy.

The components related to the aforementioned groups seem to be coherent with the confirmatory results. Indeed, the first group is mainly composed by three domains of the Connectivity index, the third by two dimensions of the Sustainability index, and the second one puts together the *Institutional* domain with many variables pertaining to the Connectivity index. The Cronbach's α are 0.96, 0.78 and 0.94, respectively.

The clustering of the statistical units returns the same results both for the confirmatory and the exploratory approach. The optimal number of clusters turns out to be equal to 6, according to the best solution of the model (1) as represented in Figure 4 by the red line.

The six clusters are pinpointed by the following countries:

1. Austria, Belgium, Denmark, Finland, Ireland, Luxembourg, Netherlands, Norway, Sweden, Switzerland, Australia, New Zealand, Singapore.
2. Brunei Darussalam, Kazakhstan, Mongolia, Russian Federation.
3. Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Greece Hungary, Latvia, Lithuania, Malta, Poland, Portugal, Romania, Slovakia, Slovenia.
4. Italy, Spain, Japan, Korea.
5. France, Germany, United Kingdom, China.

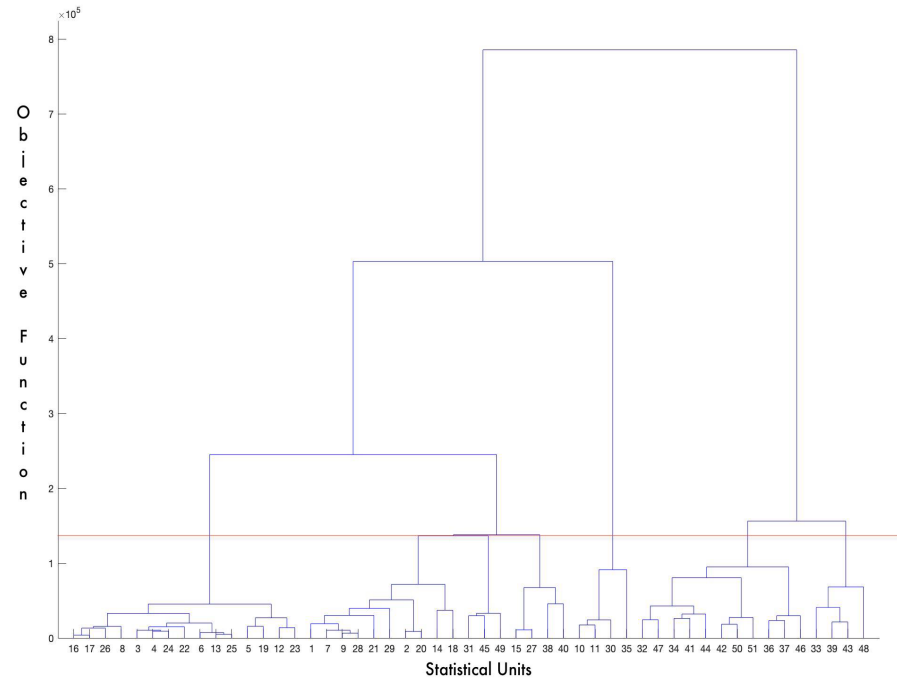


Fig. 4 Unit Clustering of the ASEM Data Set. Partition in 6 clusters of units (red line).

6. Bangladesh, Cambodia, India, Indonesia, Lao PDR, Malaysia, Myanmar, Pakistan, Philippines, Thailand, Vietnam.

The hierarchical levels from the sixth, i.e. that one with 6 clusters of countries, upwards are firstly defined by the aggregations of some of the European countries - 1 and 5 - and the Asian countries - 2 and 6. Then, the former and the remaining clusters of the European countries groups are lumped together, coherently with their geo-political distribution.

4 Conclusions

Clustering and dimensionality reduction are widely used analyses and their applications might be in several areas. Both for statistical units and for variables, the process of reduction often has a hierarchically nested shape which can be represented with a graphical configuration of a tree.

The hierarchy-shape is perfect to represent multidimensional concepts, starting from more specific ones up to the most general one, and to understand the under-

lining interconnections. A hierarchical approach permits to stop the analysis at the level the researcher considers optimal and it allows the researcher to investigate all the interconnections among items (i.e., variables and/or statistical units). In this paper, we applied the model proposed by [2] in order to get the optimal number of clusters and the optimal dimensions of the *Asia-Europe Meeting* (ASEM) data. The presence of an objective function permitted us to test a given theory and then to propose a new framework given by the study of the relations behind the data. The result is a deep study of the structure of the data and a reduced data matrix in both the dimensions (i.e., variables and/or statistical units).

References

1. Becker, W., Dominguez-Torreiro, M., Neves, A.R., Tacao Moura, C. J., Saisana, M.: Exploring ASEM Sustainable Connectivity What brings Asia and Europe together?, ISBN 978-92-79-99726-6, doi:10.2760/738153, PUBSY JRC112998 (2019)
2. Cavicchia, C., Vichi, M., Zaccaria, G.: Double Hierarchical Parsimonious Means Clustering. Unpublished manuscript
3. Cavicchia, C., Vichi, M., Zaccaria, G.: Hierarchical Disjoint Principal Component Analysis. Unpublished manuscript
4. Ferrara, C. and Martella, F. and Vichi, M.: Dimensions of Well-Being and Their Statistical Measurements. Studies in theoretical and applied statistics. 85-99 (2016)
5. Gordon, A. D.: Classification. Chapman & Hall/CRC, 2nd Edition (1999)
6. Hartigan, J. A.: Representation of Similarity Matrices by Trees. Journal of the American Statistical Association. **62:320**, 1140–1158 (1967)
7. Vichi, M., Groenen, P.K., Cavicchia, C.: Hierarchical Means Clustering. Unpublished manuscript
8. Vichi, M., Saporta, G.: Clustering and Disjoint Principal Component Analysis. Computational Statistics and Data Analysis. **53**, 3194–3208 (2009)

ICOs success drivers: a textual and statistical analysis

Fattori di successo nelle ICOs: un'analisi testuale e statistica

Paola Cerchiello and Anca Mirela Toma

Abstract Initial coin offerings (aka ICOs) represents one of the several by-product of the cryptocurrencies world. New generation start-up and existing businesses in order to avoid rigid and long money raising protocols imposed by classical channels like banks or venture capitalists, offer the inner value of their business by selling tokens, i.e. units of the chosen cryptocurrency, like a regular firm would do with and IPO. Fraudulent activities perpetrated by unscrupulous start-up happen quite often and it would be crucial to highlight in advance clear signs of illegal money raising. In this paper, we employ a statistical approach to detect what characteristics of an ICO are significantly related to fraudulent behaviours. We leverage a number of different variables like: entrepreneurial skills, number of people chatting on Telegram on the given ICO and relative sentiment, type of business, country issuing, token pre-sale price.

Abstract *Nell'ambito delle nuove Financial Technologies dette FinTech, ricoprono un ruolo importante le Initial Coin Offerings ovvero forme di raccolta di denaro a sostegno di una determinata idea di business. Le ICOs offrono in cambio un token ovvero una moneta virtuale che potr poi essere scambiata presso exchange decentralizzati. Le idee di business non sono ovviamente tutte uguali e si caratterizzano per probabilit di rischio differenti. In questo paper ci proponiamo di analizzare quali caratteristiche delle ICOs influiscano maggiormente sul relativo rischio grazie ad un'analisi statistica condotta su un dabase appositamente costruito.*

Key words: Initial coin offerings, cryptocurrencies, text analysis, fraudulent activities

Paola Cerchiello
Fintech SandLab, University of Pavia, e-mail: paola.cerchiello@unipv.it

Anca Mirela Toma
Fintech SandLab, University of Pavia, e-mail: ancamirela.toma01@universitadipavia.it

1 Introduction

Initial coin offerings (aka ICOs) are becoming more and more popular and represent an alternative strategy to raise money thanks to a new technology known as blockchain. New generation start-up and agile existing businesses, in order to avoid rigid and long money raising protocols imposed by classical channels like banks or venture capitalists, can offer the inner value of their business by selling tokens, i.e. units of a chosen cryptocurrency. Blockchain (chain of blocks) is the core technology at the basis of a cryptocurrency; it is a Distributed Ledger Technology defined as distributed, shared, encrypted database that serves as an irreversible and incorruptible repository of information (Wright, De Filippi, 2015). The number of cryptocurrencies available worldwide is close to 2100, this amount is constantly growing and if we consider market capitalization, Bitcoin is currently the largest blockchain network, followed by Ethereum, XRP, Litecoin, EOS, Bitcoin Cash, Binance Coin, Stellar, Tether and Tron, all those are struggling for the top ten market cap podium. The Bitcoin market capitalization increased from approximately 0.04 billion U.S. dollars in the first quarter of 2012 to approximately 237.62 billion U.S. dollars in the fourth quarter of 2017. It has since decreased, amounting to 66.18 billion U.S. dollars in the fourth quarter of 2018. (Coinmarketcap.com, March, 2019). The success of such decentralized technology lays on the fact that it works without the commitment and the control of a central authority: the blockchain is a Peer-to-Peer technology. The more a P2P network is distributed, scalable, autonomous and secure, the more the P2P network is valuable.

During the last year 2018, there have been 1076 ICOs, with a peak in May (145) were the ended project raised more than 21 billion US. Among those, EOS is by far the most rewarded one with 4.1 billion US followed by Telegram ICO (Pre-sale 1 & 2) that reached 1.7 billion US (Coinschedule.com, January 21, 2019).

Despite the interest arose by ICOs and the constantly growing trends, it is worth mentioning that almost half of ICOs sold in 2017 failed by February 2018 (Hankin, 2018). In fact, what should drive more attention on ICOs is the consistent presence of scam activities only devoted to raise money in a fraudulent way. It is of interest in this work to assess what factors affect the probability of success of an ICO. Relying on previous studies, Adhami et al. in 2018, based on the analysis of 253 ICOs, showed that the following characteristics contribute: the availability of the code source, the organization of a token presale and the possibility for contributors to access a specific service (or to share profits).

The main source of information about blockchains, tokens and ICOs is obviously the Web. Here we can find financial rating platforms enabling to explore the various blockchain platforms were to issue the tokens, such as the most used Ethereum, and extensive details about the ICO campaign.

On top of all the characteristics explained so far, there is a further and not yet explored point of interest: the Telegram chats. Telegram is a cloud-based instant messaging and voice over IP service developed by Telegram Messenger founded by the Russian entrepreneur Pavel Durov. As of October 2017, Telegram was by far the most popular official discussion platform for current and upcoming ICOs, with

75%+ of these projects utilizing it. These means that retrieving Telegram discussions associated to each ICOs, would produce a huge amount of textual information potentially useful for understanding the chance of success and more interestingly possible signs of scam activities.

In this paper we propose a combined approach based on classical classification models like logistic regression and random forest to highlight significant variables in distinguishing success from scam, and on text analysis. Specifically, we shall elicit from text whether some words/topic and/or a specific sentiment is expressed differently in successful/failure/scam ICOs. We have monitored and collected data (still collecting at the present date) for 120 ICOs from the beginning of 2018.

The paper is organized as follows: in section 2 we present the statistical methodology, in section 3 we describe collected data, in section 4 we illustrate results and in section 5 we report our final comments.

2 Methodology

In this paper we leverage two kinds of information: structured and unstructured ones. Regarding the former, more widely described in Section 3, we take advantage of classical statistical classification models to distinguish successful, failure and scam ICOs. Logistic regression aims to classify the dependent variable in two groups, characterized by a different status [1=scam vs 0=success or 1=success vs 0=failure] in which ICOs are classified by logistic regression, specified by the following model:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_j \beta_j x_{ij}, \quad (1)$$

where p_i is the probability of the event of interest, for ICO i , $x_i = (x_{i1}, \dots, x_{ij}, \dots, x_{iJ})$ is a vector of ICOs-specific explanatory variables, and the intercept parameter α , as well as the regression coefficients β_j , for $j = 1, \dots, J$, are to be estimated from the available data. It follows that the probability of success (or scam) can be obtained as:

$$p_i = \frac{1}{1 + \exp(\alpha + \sum_j \beta_j x_{ij})}, \quad (2)$$

Considering the textual analysis of Telegram chat we have applied a Bag of Word (BoW) approach, where a text is represented as an unordered collection of words, considering only their counts in each comment of the chat. For descriptive purposes we have used wordclouds for each and every Telegram chat according to the general content and to specific subcategories like sentiments and expressed moods.

The most critical part of the analysis relies on the sentiment classification. In general, two different approaches can be used: 1. Score dictionary based: the sentiment score is based on the number of match between predefined list of positive and negative words and terms contained in each text source (a tweet, a sentence, a whole paragraph); 2. Score classifier based: a proper statistical classifier is trained on a large enough dataset of pre-labeled examples and then used to predict the sentiment class of a new example. Insofar, we decided to focus on a dictionary based approach, adapting appropriate lists of positive and negative words relevant for ICOs topics in English language. We employ 3 vocabularies from the R package 'tidytext': AFINN, BING and NRC. (Silge, Julia & Robinson, David. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. The Journal of Open Source Software. 1. 10.21105/joss.00037.)

All three of these lexicons are based on unigrams, i.e., single words. These lexicons contain many English words and the words are assigned scores for positive/negative sentiment, and also possibly emotions like joy, anger, sadness, and so forth. The NRC lexicon categorizes words in a binary fashion (yes/no) into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The BING lexicon categorizes words in a binary fashion into positive and negative categories. The AFINN lexicon assigns words with a score that runs between -5 and 5 , with negative scores indicating negative sentiment and positive scores indicating positive sentiment.

3 Data

In this preliminary work we examine 120 ICOs starting from January 2017 to June 2018. For each project we gather information from web-based sources, mainly rating platforms such as: icobench.com, TokenData.io, ICO Drops.com, CoinDesk.com and project's websites.

In the first step of collecting process we look for general characteristics such as the name, the token symbol, start and end dates of the crowdfunding, the country of origin, financial data such as the total number of issued token, the initial price of the token, the platform used, data on the team proposing the ICO, data on the advisory board, data on the availability of the website, availability of white paper and social channels.

Some of these data, such as short and long description, and milestones are textual descriptions. Others are categorical variables, such as the country, the platform, the category (which can assume many values), and variables related to the team members (name, role, group). The remaining variables are numeric, with different degrees of discretization. Unfortunately, not all ICOs record all variables, so there are several missing data. The ICO web databases that we use are fully checked in order to minimize the missing values of one of the platforms, therefore we validate the information checking for the details on the website and on the white paper. As

a result, the complete set of reliable information comes from the matching between the website and the white paper.

A second step in the data collecting process takes into account the social channels. Those are more personal than every database, rating platform or websites, so they are a way to reach a wide range of users updating them constantly about the evolution of the project and creating a trusty environment that can finalize in a successful crowdfunding activity. In order to conduct the textual analysis, we enrich our database with the social channels data, such as the presence of a channel, the numbers of users as a proxy of the community engagement and as mentioned in the introduction the textual chat, retrieved backward till the creation of the chat. The most used social channels are Telegram, Twitter, Facebook, Bitcoinlax, Medium, while LinkedIn, Reddit and Slack are not frequently used.

With regards to the entrepreneurial dimension, we investigate the team components, pointing out that the members checked until now are almost 1000, with a median size of 7 for project. For each team member we checked general information related to the social engagement, looking for the LinkedIn channel activity (48 % of them do not have an individual page), the numbers of connections, the job position in the project and the academic background.

4 Empirical Evidence

In this section we report our main results obtained from classification analysis and textual analysis.

Regarding the former table 1 and table 2 report results respectively for logistic regression on Success/Failure (class 1 variable) and for multinomial logistic regression. The reader can see in table 3 that the only two relevant dummy variables are: the presence of a white paper and of a Telegram chat. Both present positive coefficients showing their impact on the increasing of the probability of success of an ICO. Regarding the two continuous variables, number of elements of the team and number of advisors (both appropriately standardized), are highly significant and positive suggesting that increasing people and advisors has a positive impact. In table 4 we can see results for scam ICOs, on the basis of a logistic regression modified for highly rare events as it occurs in our analysis (only 8 scam ICO out of 120 monitored). Reminding that the target variable 'class1' is labeled with 0 for scam and 1 otherwise, we can infer that both the presence of a website and of the Twitter account have a positive impact in not being a scam. In other words, the absence of these two characteristics is a driver of scam activity suspects. Considering the two continuous variables, number of components of the team and number of advisors, they have been evaluated with a non linear effect (smooth component) and similarly to previous results, we notice a positive impact. Thus, the increasing in the number of people engaged within an ICO, impacts positively on the probability of not being a scam even if not in a linear way.

Table 1 Results from Logistic regression on Success/Failure

	<i>Dependent variable:</i>
	class2
nr_team	4.522*** (1.494)
nr_adv	1.686*** (0.634)
w_paper	3.113*** (1.147)
tm	1.917** (0.955)
Constant	-2.189 (1.458)
Observations	120
Log Likelihood	-28.308
Akaike Inf. Crit.	66.616
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Table 2 Results from Gev logistic regression on Scam/non scam

	<i>Dependent variable:</i>
	class1
w_site	2.0115*** (0.490)
tw	1.230* (0.597)
s(nr_team_st)	3.973*** (smooth components)
s(nr_adv_st)	2.057*** (smooth components)
Constant	0.9894 (0.927)
Observations	120
tau	-0.25
total edf	9.03
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

5 Conclusions

In this paper we address the issue of discovering the success drivers of an ICO. Initial coin offering (aka ICO) represents one of the several by-product of the cryptocurrencies world. New generation start-up and existing businesses in order to avoid rigid and long money raising protocols imposed by classical channels like banks or venture capitalists, offer the inner value of their business by selling tokens, i.e. units of the chosen cryptocurrency, like a regular firm would do with an IPO. The investors of course hope in a value increasing of the tokens in the near future, provided a solid and valid business idea typically described by the ICO issuers in a white paper, both a descriptive and technical report of the proposed business. However, fraudulent activities perpetrated by unscrupulous start-up can happen and it would be crucial to highlight in advance clear signs of illegal money raising.

While analyzing success vs failure dynamic with a classification model is relatively easy since the incidence of the two classes is almost equal (50-50), it is much more complicated to highlight the key aspects that could witness a fraudulent activity since, in the last 3 years, only few scam events have been reported. In our sample made of 120 ICOs we have 8 scam ICOs and by fitting a logistic regression model for highly unbalance data, our preliminary results tell that both the presence of a website and of the Twitter account have a positive impact in not being a scam. In other words, the absence of these two characteristics is a driver of scam activity suspect. Considering the two continuous variables, number of components of the team and number of advisors, they have been evaluated with a non linear effect (smooth component) and we notice a positive impact. Thus, the increasing in the number of people engaged within an ICO, impacts positively on the probability of not being a scam even if not in a linear way.

References

1. Adhami, S., G. Giudici, G., Martinazzi, M.: Why Do Businesses Go Crypto? An Empirical Analysis of Initial Coin Offerings, October 20 2017, Available via SSRN.
<https://ssrn.com/abstract=3046209>
2. Calabrese, R., Marra, G., Osmetti, S.A.: Bankruptcy Prediction of Small and Medium Enterprises Using a Flexible Binary Generalized Extreme Value Model. *Journal of the Operational Research Society*, **67**(4), 604–615 (2016)
3. Coinschedule: Cryptocurrency ICO Stats 2019 available via Coinschedule.
<https://www.coinschedule.com/stats.html>. Accessed February, 2019
4. Falk, M., Haler, J., and Reiss, R.: *Laws of Small Numbers: Extremes and Rare Events*. Springer, 2010
5. Kotz, S. and Nadarajah, S.: *Extreme Value Distributions. Theory and Applications*. Imperial College Press, London, 2000
6. Mollick, E.: The dynamics of crowdfunding: an exploratory study. *J. Bus. Venturing*, **29**(1), 1–16, 2014
7. Silge, J. Robinson, D.: tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Journal of Open Source Software*, **1**, 10.21105/joss.00037, 2016
8. Subramanian H.: Decentralized Blockchain-Based Electronic Marketplaces. *Comm. ACM*, **61**, 78–84, 2018
9. Zetzsche D.A, Buckley R.P., Arner D.W. , Fhr L., : The ICO Gold Rush: It's a Scam, It's a Bubble, It's a Super Challenge for Regulators, Univ. Luxembourg Law Working Paper No. 11/2017, available via SSRN:
<https://ssrn.com/abstract=3072298>, November 2017
10. Wright, A., De Filippi, P.: Decentralized Blockchain Technology and the Rise of Lex Cryptographia, available via SSRN.
<https://ssrn.com/abstract=2580664> or <http://dx.doi.org/10.2139/ssrn.2580664>, 2015

Small area estimators with linked data

Stimatori per piccole aree nel caso di dati ottenuti attraverso il record linkage

Chambers Raymond and Fabrizi Enrico and Salvati Nicola

Abstract In Small Area Estimation data linkage can be used to combine values of the variable of interest from a national survey with values of auxiliary variables obtained from another source like a population register. Linkage errors can induce bias when fitting regression models; moreover, they can create non-representative outliers in the linked data in addition to the presence of potential representative outliers. In this paper we adopt a secondary analyst's point view, assuming limited information is available on the linkage process, and we develop small area estimators based on linear mixed and linear M-quantile models to accommodate linked data containing a mix of both types of outliers. A simulation exercise is presented to illustrate the main results.

Abstract Nella stima per piccole aree il data linkage può essere utilizzato per combinare i valori della variabile di interesse ottenuti da un'indagine campionaria con i valori delle variabili ausiliarie ottenuti da un'altra fonte, tipicamente un registro. Gli errori di linkage possono indurre distorsione nella stima dei parametri dei modelli di regressione; inoltre, possono creare valori anomali non rappresentativi accanto ai valori anomali rappresentativi potenzialmente presenti nei dati. In questa ricerca adottiamo la prospettiva di un analista secondario, supponendo che siano disponibili informazioni limitate sul processo di linkage e sviluppiamo estimatori per piccole aree basati su modelli lineari misti e M-quantilici per trattare dati linkati contenenti un mix di entrambi i tipi di valori anomali. Viene presentato un esercizio di simulazione per illustrare i principali risultati.

Key words: Exchangeable linkage error; Finite population inference; Linear mixed models; Mean Squared Error estimation; Robust estimation.

Chambers Raymond
University of Wollongong, e-mail: ray@uow.edu.au

Enrico Fabrizi
DISES, Università Cattolica del S. Cuore, Via Emilia Parmense 84, Piacenza e-mail: enrico.fabrizi@unicatt.it and Salvati Nicola
Università di Pisa, Via Ridolfi 10, 56124 Pisa

1 Introduction

A small area estimation problem arises when estimates for a collection of subsets (often called areas) of a target population are needed, but area-specific samples are too small for ordinary estimators of population descriptive quantities to reach adequate precision. Small area estimation methods complement survey data on the target variable y , with area specific auxiliary information. A standard setting is where it is reasonable to assume that the value y_{ij} for unit j in area i is related to a known vector of covariates \mathbf{x}_{ij} by means of a regression model. These \mathbf{x}_{ij} values, assumed to be known for both the survey sample and the rest of the population, are then used to predict the area parameter of interest.

Data integration is fast becoming an intrinsic part of Official Statistics, in large part due to the increasing availability of administrative registers and other population data sources. Here we focus on the situation where the y_{ij} values are measured in a sample survey but where the \mathbf{x}_{ij} are extracted from a population register and then linked to the sampled units. In many cases an error-free unique identifier is not available and we need to allow for record linkage errors.

We assume the point of view of a secondary data analyst with limited access to information related to data linkage process, so our attention will be devoted to the analysis of linked data rather than to the process of record linkage. Overlooking linkage errors when analysing linked data can lead to biased estimates even if most records are correctly linked. Bias correction methods when fitting linear regression models to linked data are discussed in Scheuren and Winkler (1993), Scheuren and Winkler (1997), Lahiri and Larsen (2005), Chambers (2009), Kim and Chambers (2012) and Han and Lahiri (2018). The impact of linkage errors on linear mixed models, which are often used in small area estimation, has received comparatively less attention (Samart and Chambers, 2014). More specifically, we are aware of just one other article (Briscolini *et al.*, 2018) where linear mixed models are used with linked data for small area estimation.

However, there is another aspect to linkage errors that seems to have attracted much less attention. This is when linkage errors generate artificial outliers in the linked data set. Let y_{ij}^* denote the linked value corresponding to y_{ij} . Such an outlier can then be generated when there is linkage error, and the residual associated with the correctly linked pair $(y_{ij}, \mathbf{x}_{ij})$ is small, but the residual associated with the incorrectly linked pair $(y_{ij}^*, \mathbf{x}_{ij})$ is large. This can happen when the variables used in the matching process (such as names, addresses, identification codes) are independent of those used as regressors.

Using the distinction between representative and non-representative outliers introduced by Chambers (1986), these artificial outliers are non-representative, and so are fundamentally different from outliers associated with the correctly linked population units, which are representative. The problem is that it is not possible to tell *a priori* whether an outlier is induced by linkage error (and so is non-representative) or is representative.

Two questions immediately arise. The first is whether well-known outlier robust methods for small area estimation can adequately deal with the mix of representative

and non-representative outliers that can potentially occur in a linked data situation. In the rest of this paper we introduce basic notation and concepts (section 2), the modification of linear mixed models and M-quantile linear regression needed to allow for linkage errors (section 3) and the results from a simulation exercise (section 4). Technical results are omitted due to lack of space.

2 Background and notation

We assume that we have two registers covering the same population U of size N and that can in principle be linked one-to-one. The first register, labelled Y is the one from which the target variable is sampled, while the second, X contains the auxiliary variables.

We assume that each register is partitioned into Q non-overlapping blocks such that linkage errors are homogeneous within a block. Moreover we assume that linkage errors are possible within blocks but not across blocks. For the purposes of small area estimation U can be partitioned into D non-overlapping areas, and that the linking is carried out within an area, so two population units from different areas cannot be erroneously matched. Cross-classifying U by the area and block indicators, we then define U_{iq} to be the subset of N_{iq} population units that make up the segment of area i nested within block q , with $i = 1, \dots, D$ and $q = 1, \dots, Q$. We use \mathbf{x}_{iqj} and \mathbf{y}_{iqk} to denote individual population values from the two registers associated with this iq cell.

Let's suppose a sample is drawn with a non-informative sampling design. Let s_{iq} denote the set corresponding to the n_{iq} population indexes of the sample units in small area i and block q , with $n = \sum_{i=1}^D \sum_{q=1}^Q n_{iq}$. Let \mathbf{y}_{iq} denote the N_{iq} vector of values for y_{iqk} in U_{iq} , with \mathbf{X}_{iq} denoting the $N_{iq} \times p$ matrix with rows defined by the \mathbf{x}_{iqj} values of the corresponding population units. We characterise the relationship between \mathbf{y}_{iq} and \mathbf{y}_{iq}^* via a latent random permutation matrix $\mathbf{A}_{iq} = [a_{jk}^{iq}]$ of order N_{iq} . That is, we put

$$\mathbf{y}_{iq}^* = \mathbf{A}_{iq} \mathbf{y}_{iq}. \quad (1)$$

We assume that information available to the secondary data analyst reduces to parameters involved in this simple exchangeable linkage error (ELE) model:

$$Pr(\text{correct linkage}) = Pr(a_{jj}^{iq} = 1) = \lambda_q \quad (2)$$

$$Pr(\text{incorrect linkage}) = Pr(a_{jk}^{iq} = 1) = \gamma_q = \frac{1 - \lambda_q}{N_{iq} - 1}, \quad (3)$$

with $j, k = 1, \dots, N_{iq}$. We shall assume that the values of λ_q are known or can be accurately estimated from the information in the linkage paradata. Unless stated otherwise, from now on we therefore condition our analysis on these values of λ_q . Assuming that the linkage is non-informative for the distribution of \mathbf{y}_{iq} given \mathbf{X}_{iq}

we have a consequence that $E_A(\mathbf{A}_{iq}|\mathbf{X}_{iq}) = \mathbf{T}_{iq} = (\lambda_q - \gamma_q)\mathbf{I}_{N_{iq}} + \gamma_q\mathbf{1}_{N_{iq}}\mathbf{1}'_{N_{iq}}$ where $\mathbf{I}_{N_{iq}}$ denotes the identity matrix of order N_{iq} , $\mathbf{1}_{N_{iq}}$ denotes a vector of ones of length N_{iq} and $E_A(\cdot)$ denotes expectation with respect to the linkage error model. It immediately follows that we can write,

$$E_ME_A(\mathbf{A}_{iq}\mathbf{y}_{iq}|\mathbf{X}_{iq}) = E_A(\mathbf{A}_{iq}|\mathbf{X}_{iq})E_M(\mathbf{y}_{iq}|\mathbf{X}_{iq}) = \mathbf{T}_{iq}E_M(\mathbf{y}_{iq}|\mathbf{X}_{iq}). \quad (4)$$

Here $E_M(\cdot)$ denotes expectation with respect to the model for \mathbf{y}_{iq} given \mathbf{X}_{iq} .

We note that in (4) we have two sources of randomness and that moments are taken with respect to the joint probability space defined by the linkage process and the assumed population model. Focusing on the relationship at the sample level let \mathbf{A}_{siq} and \mathbf{A}_{riq} contain the rows of \mathbf{A}_{iq} corresponding to sampled and non-sampled units, respectively. Then \mathbf{y}_{siq}^* , \mathbf{X}_{iq} are observed, i.e. available to the analyst, while \mathbf{y}_{siq} is not observed, where $\mathbf{y}_{siq}^* = \mathbf{A}_{siq}\mathbf{y}_{iq}$. The matrix \mathbf{A}_{siq} is not observable, but under the ELE assumptions (2) and (3) we have that $E_A(\mathbf{A}_{siq}|\mathbf{X}_{iq}) = \mathbf{T}_{siq}$. As we condition on λ_q \mathbf{T}_{siq} can be treated as known. We assume that the sampled rows and the column means of \mathbf{X}_{iq} are known. As a consequence, the matrix $\mathbf{X}_{siq}^* = E_A(\mathbf{A}_{siq}\mathbf{X}_{iq}|\mathbf{X}_{iq}) = \mathbf{T}_{siq}\mathbf{X}_{iq}$ can be treated as known.

3 Small area models for linked data

3.1 Linear mixed models

Linear mixed models for population unit data are widely used for SAE. A general specification for a unit level linear mixed model used in SAE is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (5)$$

where \mathbf{y} and \mathbf{X} denote the population level vector of response variable and matrix of covariates, respectively; $\mathbf{u} = (\mathbf{u}'_1, \dots, \mathbf{u}'_D)$ is a vector of dimension Dm made up of D independent realizations $\{\mathbf{u}_i; i = 1, \dots, D\}$ of a m -dimensional random area effect with $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u)$ and $\mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_e)$ is the $N \times 1$ vector of individual errors. Since the random effects \mathbf{u} and the individual errors \mathbf{e} are independent, the covariance matrix of \mathbf{y} is $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_e + \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}'$. Here D is the total number of small areas that make up the population and m is the dimension of \mathbf{z}_{ij} so that \mathbf{Z} is an $N \times Dm$ matrix of fixed known constants that do not vary within an area. We assume that the covariance matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ are defined in terms of a lower dimensional set of parameters $\boldsymbol{\delta} = (\delta_1, \dots, \delta_K)$, which are typically referred to as the variance components of model (5), whereas the vector $\boldsymbol{\beta}$ stands for the $p \times 1$ vector of regression coefficients. Provided that it is reasonable to assume that the distribution of the unit level residuals in \mathbf{e} remains the same from block to block, then at the U_{iq} sub-population level, (5) can be written as:

$$\mathbf{y}_{iq} = \mathbf{X}_{iq}\beta + \mathbf{Z}_{iq}\mathbf{u}_i + \mathbf{e}_{iq}, \quad (6)$$

where \mathbf{Z}_{iq} is the $N_{iq} \times m$ incidence matrix defined by the rows of \mathbf{Z} corresponding to area i units in block q .

Now suppose the sample selection process is non-informative, we can use (2-3) to write down a model for the linked sample values \mathbf{y}_{siq}^* that takes into account the linkage error process:

$$\mathbf{y}_{siq}^* = \mathbf{A}_{siq}\mathbf{y}_{iq} = \mathbf{A}_{siq}\mathbf{X}_{iq}\beta + \mathbf{Z}_{siq}\mathbf{u}_i + \mathbf{e}_{siq}^*, \quad (7)$$

where $E_M(\mathbf{e}_{siq}^*) = \mathbf{0}$ and $V_M(\mathbf{e}_{siq}^*) = \boldsymbol{\Sigma}_{seiq}$, i.e. the sampled rows of the area i by block q component of $\boldsymbol{\Sigma}_e$. Also, since the values contained in the columns of \mathbf{Z}_{iq} do not change within an area, and because we assume that matching across areas is impossible, it follows that $\mathbf{A}_{siq}\mathbf{Z}_{iq} = \mathbf{Z}_{siq}$, i.e. linkage errors have no impact on the sampled rows of \mathbf{Z}_{iq} . We then have that $E_{\mathbf{A},M}(\mathbf{y}_{siq}^*|\mathbf{X}_{iq}) = \mathbf{X}_{siq}^*\beta$, $V_{\mathbf{A},M}(\mathbf{y}_{siq}^*|\mathbf{X}_{iq}) = \boldsymbol{\Sigma}_{siq} = \mathbf{Z}_{siq}\boldsymbol{\Sigma}_{\mathbf{u}_i}\mathbf{Z}_{siq}' + \boldsymbol{\Sigma}_{seiq} + \mathbf{V}_{siq}$, where $E_{\mathbf{A},M}$ and $V_{\mathbf{A},M}$ are the joint expectation and variance with respect to the linkage error model and the linear mixed model respectively. Here $\boldsymbol{\Sigma}_{\mathbf{u}_i}$ is a $m \times m$ matrix, corresponding to the i th diagonal block of $\boldsymbol{\Sigma}_{\mathbf{u}}$, and $\mathbf{V}_{siq} = V_{\mathbf{A}}(\mathbf{A}_{siq}\mathbf{X}_{iq}\beta)$. An exact expression for \mathbf{V}_{siq} is unavailable. However, using the arguments set out in Chambers (2009), we can write down the approximation

$$\mathbf{V}_{siq} \approx \text{diag}\{\mathbf{v}_{siq}\} = \text{diag}\left((1 - \lambda_q)(\lambda_q(f_{iqj} - \bar{f}_{siq})^2 + \bar{f}_{siq}^{(2)} - \bar{f}_{siq}^2)\right), \quad j = 1, \dots, n_{iq}, \quad (8)$$

where $\mathbf{f}_{siq} = \{f_{iqj}\} = \mathbf{x}_{ijq}'\beta$ and \bar{f}_{siq} , $\bar{f}_{siq}^{(2)}$ denote the block q averages of the components of \mathbf{f}_{siq} and their squares, respectively.

In this research we analyze this model and derive the EBLUP (\hat{Y}_i) and the REBLUP predictor (\hat{Y}_i) for \bar{Y}_i in the case of linked data. For the REBLUP we extend the methodology of Sinha and Rao (2009). Moreover we obtained analytical expression for the approximated MSE of these predictors.

3.2 Linear M-quantile models

M-quantile regression models were first suggested for small area estimation by Chambers and Tzavidis (2006). See Bianchi *et al.* (2018) for a recent review of subsequent applications and theoretical extensions. In the linear case, M-quantile regression leads to a family of hyperplanes indexed by a real number $\tau \in (0, 1)$ representing the order of the M-quantile of interest, i.e. $MQ(\tau|\mathbf{x}_{ij}) = \mathbf{x}_{ij}'\beta_\tau$.

For specified τ and influence function ψ (with $\psi_\tau = d\rho_\tau(u)/du$) an estimate $\hat{\beta}_\tau$ of the vector of regression parameters β_τ may be obtained as the solution to the normal equations,

$$\sum_{i=1}^D \sum_{j=1}^{n_i} \psi_\tau \left(\frac{y_{ij} - \mathbf{x}_{ij}'\hat{\beta}_\tau}{\sigma_\tau} \right) \mathbf{x}_{ij} = \mathbf{0}, \quad (9)$$

where σ_τ is a scale parameter that characterizes the spread of the distribution of the residuals $y_{ij} - \mathbf{x}'_{ij}\beta_\tau$. Following standard practice in robust M-regression, this scale parameter can be estimated by $\hat{\sigma}_\tau = \text{median}\{|y_{ij} - \mathbf{x}'_{ij}\hat{\beta}_\tau|\}/0.6745$.

Following the approach of Chambers (2009), we therefore modify the M-quantile normal equations (9) to take account of the linkage error structure, using the notation introduced in Section 2. This leads to the modified M-quantile normal equations,

$$\sum_{i=1}^D \sum_{q \in i} \mathbf{X}_{siq}^* \mathbf{R}_{siq\tau}^{-1/2} \psi_\tau \left\{ \mathbf{R}_{siq\tau}^{-1/2} (\mathbf{y}_{siq}^* - \mathbf{X}_{siq\tau}^* \beta_\tau^*) \right\} = \mathbf{0}, \quad (10)$$

where $\mathbf{R}_{siq\tau} = \text{diag} \left(\sigma_\tau^{*2} + (1 - \lambda_q)(\lambda_q(f_{iqj\tau} - \bar{f}_{siq\tau})^2 + \bar{f}_{siq\tau}^{(2)} - \bar{f}_{siq\tau}^2) \right)$, $\mathbf{f}_{siq\tau} = \{f_{iqj\tau}\} = \mathbf{x}'_{iqj}\beta_\tau^*$, and $\bar{f}_{siq\tau}$, $\bar{f}_{siq\tau}^2$ denote the block q averages of the components of $\mathbf{f}_{siq\tau}$ and their squares respectively. Note that σ_τ^* here is the scale coefficient associated with the skewed residuals from the M-quantile regression line of order τ . In this research we then developed the predictor \hat{Y} based on the estimating equations (10) and obtain also an approximation to itm MSE.

4 Simulation exercise

We used model-based simulations of the various linked data based small area predictors previously described. The synthetic populations underpinning these simulations are based on those used by Chambers *et al.* (2014) with some modifications.

Values for y are generated from the equation $y_{ij} = 100 + 5x_{ij} + u_i + \varepsilon_{ij}$ $j = 1, \dots, N_i$, $i = 1, \dots, D$; values for x are generated independently from a lognormal distribution with a mean of 1.0 and a standard deviation of 0.5 on the log-scale. The population is divided in 40 areas ($i = 1, \dots, D = 40$) each of size $N_i = 100$. The random components u_i and ε_{ij} are generated independently according to two scenarios:

- (0,0) $u \sim N(0, 3)$ and $\varepsilon \sim N(0, 6)$. In this scenario there are artificial outliers caused by linkage errors.
- (e,u) $u \sim N(0, 3)$ for areas 1–36, $u \sim N(0, 20)$ for areas 37–40 and $\varepsilon \sim \delta N(0, 6) + (1 - \delta)N(0, 150)$ where δ is an independently generated Bernoulli random variable with $Pr(\delta = 1) = 0.97$, i.e. the individual effects are independent draws from a mixture of two normal distributions, with 97% on average drawn from a ‘well-behaved’ $N(0, 6)$ distribution and 3% on average drawn from an outlier $N(0, 150)$ distribution.

Linked data pairs $(x_{ij}; y_{ij}^*)$ are then generated using the exchangeable linkage errors model (1) with correct linkage probabilities $\lambda_q = 1.0, 0.9, 0.6$ and 0.4 for blocks 1, 2, 3 and 4, respectively. In each area there are 25 units for each block, assigned randomly. In scenario (e, u) there are both artificial and real outliers. Samples of size $n_i = 5$ are selected by simple random sampling without replacement within each

area. Results comparing the performances of various estimators are displayed in table 1. Estimators \hat{Y}_i^{EBLUP} , \hat{Y}_i^{REBLUP} and \hat{Y}_i^{MQ} are the ordinary small area predictors that overlook record linkage. Estimators are compared in terms of median percentage of relative bias and the median percentage of Relative Root MSE (RRMSE).

Predictor	Results (%) for the following scenarios:			
	Relative bias		RRMSE (EFF)	
	(0,0)	(e,u)	(0,0)	(e,u)
\hat{Y}_i^{EBLUP}	0.00	-0.01	1.37 (100.0)	1.42 (100.0)
\hat{Y}_i^{*EBLUP}	0.03	0.01	1.26 (91.9)	1.28 (90.2)
\hat{Y}_i^{REBLUP}	-0.06	-0.06	1.16 (84.3)	1.20 (83.7)
$\hat{Y}_i^{*REBLUP}$	-0.09	-0.09	1.14 (82.7)	1.17 (81.2)
\hat{Y}_i^{MQ}	-0.19	-0.17	1.31 (94.8)	1.34 (92.9)
\hat{Y}_i^{*MQ}	-0.04	-0.05	1.12 (81.4)	1.17 (80.8)

Table 1 Model-based simulation results: median values of the percentage of relative bias and RRMSE of predictors of small area means with $n_i = 5$. In parenthesis the values of the efficiency (EFF).

The results set out in table 1 are in line with our expectations regarding the behaviour of the corrected predictors based on the linked data. They show smaller bias and higher effinecy than the traditional small area estimators in both scenarios. The performances of the MSE estimators for the EBLUP, REBLUP and M-quantile-based predictors are evaluated in Table 2. Here we are mainly interested in the performances of the MSE estimators for the corrected predictors based on the linked data.

Predictor	Results (%) for the following scenarios:			
	Relative bias		RRMSE	
	(0,0)	(e,u)	(0,0)	(e,u)
$\widehat{MSE}(\hat{Y}_i^{EBLUP})$	-3.9	2.8	22.5	32.1
$\widehat{MSE}_{A,M}(\hat{Y}_i^{*EBLUP})$	0.7	14.1	21.5	35.3
$\widehat{MSE}(\hat{Y}_i^{REBLUP})$	9.1	6.1	52.8	44.4
$\widehat{MSE}_{A,M,u}(\hat{Y}_i^{*REBLUP})$	-0.4	19.2	51.1	63.0
$\widehat{MSE}(\hat{Y}_i^{MQ})$	-10.3	-12.6	57.6	56.9
$\widehat{MSE}_{A,M}(\hat{Y}_i^{*MQ})$	-2.7	-4.7	43.1	44.1

Table 2 Median values of percentage of relative bias and RRMSE of RMSE estimators in model-based simulation experiments.

From table 2 We see that the MSE estimator for \hat{Y}_i^{*EBLUP} tends to be somewhat biased low under the (0,0) scenario. The MSE estimator for \hat{Y}_i^{*MQ} is less biased

than the corresponding estimator for $\hat{Y}_i^{*REBLUP}$ for this scenario. In general, the proposed MSE estimators work well under linkage errors in scenario (0,0). Under the (e,u) scenario the MSE estimators of \hat{Y}_i^{*EBLUP} and the MSE estimator of $\hat{Y}_i^{*REBLUP}$ all tend to overestimate the actual MSE, whereas the MSE estimator for \hat{Y}_i^{*MQ} is slightly negatively biased. We also see that the MSE estimators of \hat{Y}_i^{EBLUP} , \hat{Y}_i^{*EBLUP} and $\hat{Y}_i^{**EBLUP}$ are generally more stable than those for the REBLUP and M-quantile-based predictors.

References

- Bianchi, A., Fabrizi, E., Salvati, N., and Tzavidis, N. (2018). Estimation and testing in m-quantile regression with applications to small area estimation. *International Statistical Review*, f86, 541–570.
- Briscolini, D., Consiglio, L. D., Liseo, B., Tancredi, A., and Tuoto, T. (2018). New methods for small area estimation with linkage uncertainty. *International Journal of Approximate Reasoning*, f94, 30–42.
- Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, f81, 1063–1069.
- Chambers, R. (2009). Regression analysis of probability-linked data. Technical report, Official Statistics Research, Statistics New Zealand, Downloaded from <http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm>.
- Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, f93 (2), 255–268.
- Chambers, R., Chandra, H., Salvati, N., and Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B*, f76 (1), 47–69.
- Han, Y. and Lahiri, P. (2018). Statistical analysis with linked data. *International Statistical Review*, doi:10.1111/insr.12295, 1–19.
- Kim, G. and Chambers, R. (2012). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, f56, 2756–2770.
- Lahiri, P. and Larsen, M. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, f100 (469), 222–230.
- Samart, K. and Chambers, R. (2014). Linear regression with nested errors using probability-linked data. *Australian and New Zealand Journal of Statistics*, f56 (1), 27–46.
- Scheuren, F. and Winkler, W. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, f19, 39–58.
- Scheuren, F. and Winkler, W. (1997). Regression analysis of data files that are computer matched - part ii. *Survey Methodology*, f23, 157–165.
- Sinha, S. K. and Rao, J. N. K. (2009). Robust small area estimation. *The Canadian Journal of Statistics*, f37 (3), 381–399.

Optimal Portfolio Selection via network theory in banking and insurance sector

Gian Paolo Clemente, Rosanna Grassi and Asmerilda Hitaj

Abstract This paper focuses on network portfolio selection approach based on different estimation methods for the covariance matrix. In particular the sample and shrinkage toward the constant correlation estimators are tested. A case study based on asset belonging to banking and insurance sector is developed.

Key words: Portfolio selection, Networks, Dependence, Interconnectedness, Banking and Insurance sector

1 Introduction

Modern portfolio theory (e.g. [15]) allocates the wealth across a set of assets considering only first and second moments. The out-of sample performance of this model is often affected by large estimation errors on the mean and covariance matrix. To overcome this issue, shrinkage approaches are introduced (see among others [13, 16]). Indeed, the use of shrinkage estimators for moments and co-moments often improves the out of sample performance, as reported in [10, 11]. The authors in [6] tackle the asset allocation problem by considering the portfolio as a network. In particular, they catch how much a node is embedded in the system, by adapting to this context the clustering coefficient, a specific network index (see

Gian Paolo Clemente

Universit Cattolica del Sacro Cuore, Department of Mathematics, Finance and Econometrics, Milano e-mail: gianpaolo.clemente@unicatt.it

Rosanna Grassi

University of Milano - Bicocca, Department of Statistics and Quantitative Methods, Milano e-mail: rosanna.grassi@unimib.it

Asmerilda Hitaj

University of Milano - Bicocca, Department of Statistics and Quantitative Methods, Milano e-mail: asmerilda.hitaj1@unimib.it

[1, 5, 8, 17, 22, 23]), meaningful in financial literature to assess systemic risk [3, 19, 21]). In constructing the dependence structure of the portfolio network, various dependence measures are tested, namely, the Pearson correlation, Kendall correlation and lower tail dependence. These measures are estimated using the sample approach. The results obtained in [6] showed that, independently from the length of the rolling window and from the dependence structure used, the network-based portfolio leads to better out-of-sample performance compared with the classical approach. In this paper we move one step further and make a first attempt in analyzing the effect of the estimation method on the network portfolio selection model. We compare *classical global minimum variance portfolio* GMV approach and a *network* approach based on linear correlation (NB). For each model we consider two different estimation methods for the covariance matrix: *sample* and *shrinkage*. The impact of the estimation method is tested on a real portfolio, characterized by the worldwide largest banks and insurance companies and by comparing several in-sample and out-of-sample measures. Main results show that, as expected, the use of shrinkage estimators lead to a higher out of-sample performance in the GMV approach. The network-based approach is instead more robust being only slightly affected by the estimation method of the covariance matrix.

The remainder of the paper is organized as follows. In Section 2 the two estimation methods are briefly explained. Section 3 recalls the approach of portfolio selection via network theory. Section 4 presents the empirical analysis and Section 5 draws some conclusions.

2 Classical portfolio selection and estimation methods for covariance matrix

The GMV portfolio is obtained by solving the following optimization problem:

$$\begin{cases} \min_{\mathbf{x}} & \mathbf{x}^T \mathbf{\Sigma} \mathbf{x} \\ & \mathbf{e}^T \mathbf{x} = 1 \\ & 0 \leq x_i \leq 1, \quad i = 1, \dots, N \end{cases}, \quad (1)$$

where $\mathbf{x} = [x_i]_{i=1, \dots, N}$ is the vector of portfolio weights and $\mathbf{\Sigma}$ is the covariance matrix. The first equation stands for the budget constraint and the second constraint excludes the possibility of short selling. In order to solve the investment problem we need to estimate the covariance matrix between assets in a given time interval. A common way is the *sample approach*, where each component $\hat{\sigma}_{i,j}$ of the matrix is estimated by using classical unbiased estimator.

It is well known that the sample estimator of historical returns is likely to generate high sampling error. For this reason several methods have been introduced to improve the covariance matrix estimation. The idea is to impose some structure on the covariance matrix that limits the number of parameters leading to a reduction of the

sampling error at the cost of specification error. In this work we consider only the shrinkage toward the constant correlation (CC) method (see [7, 13]). The idea of the CC approach is to estimate the covariance based on the fact that the correlation is assumed constant for each pair of assets, and it is given by the average of all the sample correlation coefficients (see [7]). The covariance between two assets is then computed as $\sigma_{i,j}^{CC} = \hat{\sigma}_i \hat{\sigma}_j \frac{1}{N(N-1)} \sum_{\substack{j=1 \\ i \neq j}}^N (\hat{\rho}_{i,j})$, where $\hat{\rho}_{i,j}$ is the sample correlation

between asset i and j .

This approach resizes the problem, as only one correlation coefficient and N standard deviations have to be estimated. The Σ^{CC} covariance matrix, constructed by using previous formula, is characterized by a lower estimation risk due to the assumed structure, nevertheless involves some misspecification in the artificial structure imposed by this estimator. In the attempt to find a trade-off between the sample risk and the model risk, the authors in [13] introduce the asymptotically optimal linear combination of the sample estimator and the structured estimator (in our case is the CC one) in the context of the covariance matrix, with the weight given by the optimal shrinkage intensity κ^1 . Therefore, the shrinkage toward CC covariance matrix is given by:

$$\sigma_{i,j}^{shrink} = \kappa \sigma_{i,j}^{CC} + (1 - \kappa) \hat{\sigma}_{i,j}. \quad (2)$$

We investigate how the use of shrinkage estimators perform in case of network portfolio selection approach.

3 Optimal portfolio via network theory

The financial market is represented as a network ([6]), where assets are nodes and weighted edges accounts for the correlation between returns. The developed investment strategy benefits from the knowledge of such market dependency structure. Indeed, unlike the GMV model, we provide an objective function that takes into account the interconnectedness of the system, not only the pairwise correlations.

Let $G = (V, E)$ be a network with the set V of nodes and the set E of edges between nodes. The edge (i, j) connects a pair of nodes i and j . If $(j, i) \in E$ whenever $(i, j) \in E$, the network is undirected. A network is complete if every pair of vertices is connected by an edge. We denote with k_i the degree of the node i ($i = 1, \dots, n$) and \mathbf{A} its adjacency matrix, whose elements are $a_{ij} = 1$ whenever $(i, j) \in E$ and 0 otherwise. A network is weighted if a weight $w_{ij} \in \mathbb{R}$ is associated to each edge (i, j) . In this case, both adjacency relationships between vertices of G and weights on the edges are described by a non negative, real N -square matrix \mathbf{W} (the weighted adjacency matrix).

We consider here a weighted, complete and undirected network G that represents the correlations structure, being all assets correlated, with edges' weight equal to

¹ For further information on how is estimated the shrinkage intensity κ see [13].

the linear correlation between them². We highlight that, in constructing the portfolio network, we consider either the sample or the shrinkage estimators of the covariance matrix in NB-sample or NB-shrinkage approach, respectively. The following steps and the formulation of the problem are the same in both cases.

To capture the level of interconnectivity of an asset with the entire system, we make use of the clustering coefficient. Being complete the correlation graph, the classic clustering coefficient and its variants (see [23, 4, 8, 5]) are not computable and the measure have to be adapted to our framework. We follow a procedure in line with [17], modifying the adjacency matrix fixing a threshold $s \in [-1, 1]$ and defining \mathbf{A}_s , whose elements a_{ij}^s are equal to 1 if $w_{ij} \geq s$ and 0 otherwise. This is the adjacency matrix describing the existing links in the network with weights w_{ij} at, or above the threshold s . The idea is to capture the mean cluster prevalence of the network looking at a zoom-out level where only the strongest edges (i.e. greater than a given threshold) are visible.

Hence, we compute the clustering coefficient proposed in [23] and then we repeat the process, varying the threshold s . The clustering coefficient C_i for a node i corresponding to the graph is the average of $C_i(\mathbf{A}_s)$ overall $s \in [-1, 1]$:

$$C_i = \int_{-1}^1 C_i(\mathbf{A}_s) ds \quad (3)$$

Since $0 \leq C_i \leq 1$, C_i is well-defined. Now, we define the N -square matrix \mathbf{C} , of entries $c_{ij} = C_i C_j$ if $i \neq j$ and 1 otherwise.

Since this matrix accounts for the interconnection level of each couple with the system, we can interpret it as an interconnectedness matrix. We define the matrix $\mathbf{H} = \mathbf{\Delta}^T \mathbf{C} \mathbf{\Delta}$ where $\mathbf{\Delta} = \text{diag}(s_i)$ is a diagonal matrix with diagonal entries $s_i = \frac{\hat{\sigma}_i}{\sqrt{\sum_{i=1}^N \hat{\sigma}_i^2}}$ ³. The network based portfolio model is obtained by solving the optimization problem defined in (1) replacing the covariance matrix $\mathbf{\Sigma}$ with \mathbf{H} ⁴.

The main difference between the classical and the network portfolio selection is due to the use of the interconnectedness matrix in order to consider how much each couple of assets is related to the system. In particular, being \mathbf{C} dependent on a network-based measure of systemic risk (i.e. the clustering coefficient), we are implicitly including a measure of the state of stress of the financial system in the time period.

² An ultrametric distance can be associated with the correlation coefficient to assure that weights take positive values (see [9, 14, 20]). In our case, this transformation does not affect the results in terms of optimal portfolio.

³ The element s_i considers the contribute of the standard deviation of the returns i with respect to the total standard deviation, computed in case of independence.

⁴ Observe that the matrix \mathbf{H} is positive semidefinite (see [6]). In addition, the objective function is continuous in a feasible compact set, then at least one solution of the problem exists.

4 Dataset description and Results

In order to assess the effectiveness of the four models under analysis we perform some empirical applications. The investment universe of the considered portfolio is composed by 266 among largest banks and insurance companies in the world⁵. The dataset contains daily returns of 120 insurers and 144 banks in the time-period ranging from January 2001 to the end of 2017⁶.

The obtained portfolios are analyzed and compared in an in-sample and out-of-sample perspective. In an in-sample perspective we consider the modified Herfindhal index as a measure of diversification and the transaction costs. In an out-of-sample perspective the first four moments, the Sharpe Ratio (SR), the Omega Ratio (OR) and the Information Ratio (IR) are used⁷. All these aspects are investigated through a rolling window methodology, which is characterized by an in-sample period of length n and an out-of-sample period of length h .

In this empirical analysis we consider two different rolling-window strategies; namely, 6 months in-sample and 1 month out-of-sample and two years in-sample and 1 month out-of-sample. Figure 1 (left-hand side) depicts the correlation network obtained in the last window that covers the period December 2015–November 2017. As previously pointed out, each node represents a firm (bank or insurer) and the weighted edge (i, j) measures the correlation between firms i and j , based on the sample estimation. As described in Section 3, correlation has been used to assess the interconnections structure of each asset by means of a clustering coefficient. Hence, the (modified) problem (1) is solved. We report in Figure 1, right-hand side, the optimal solution for the same window w as in the left-hand side. In this network representation, size of bullets is instead proportional to allocated weight. We observe that the initial endowment is invested in only 26 firms, 10 banks and 16 insurance companies. However, approximately 94% of the total amount is invested in insurers, that, in this time period, are characterized on average by both a lower volatility and a lower clustering coefficient. We can remark the case of two insurers (Nationwide Mutual Insurance Company and One America), that are characterized by the lowest standard deviations between the firms and by a high proportion of negative pairwise correlations (for instance, approximately 90% of correlations between Nationwide Mutual and other firms is lower than zero). As expected, the optimal portfolio is based on a high proportion of the initial endowment invested in these two firms (54% and 17% respectively).

Now, we compare the four models under analysis using the rolling window procedure, where the correlation matrix has been estimated via sample or shrinkage method, respectively. Portfolio diversification and the portfolio turnover results are

⁵ The greatest firms by market capitalization in the banking and insurance sector are considered.

⁶ Data have been downloaded from Bloomberg [2]

⁷ These measures are well known in the financial literature and their definition, for space constraints, is not reported. The empirical protocol in this paper is the same as in [6], where interested reader can find a complete definition of all these measure.

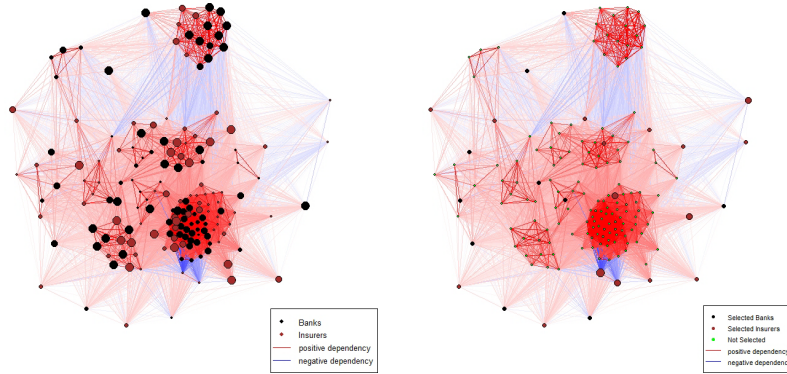


Fig. 1: On the left-hand side, we report the Pearson Correlation Network computed by using returns of Banks and Insurers dataset (based on sample estimation). These Figures refers to the last window that goes from the 1st of December 2015 to the end of November 2017. Bullets size is proportional to the standard deviation of each firm. Edges opacity is proportional to weights (i.e., intensity of correlation). On the right-hand side, the optimal portfolio for the same period is displayed. Bullets size is proportional to allocated weight. Edges opacity is proportional to weights.

reported in Figures 3 and 4⁸. As already noticed in [6], network-based approach is characterized on average by a lower turnover, namely, allocated weights have been rebalanced less than GMV approach. Main results are reported in Figure 2, and for both rolling windows in Table 1. Concerning the performance we observe, in Figure 2, that although the network-based approach leads to almost the same amount at the end of the period, it always shows a performance higher than GMV-sample and GMV-shrinkage, with a significant over-performance in the out-of-sample windows starting in 2006-2007 and since the end of 2009 to the end of 2014. We also notice that the GMV-shrinkage leads to a higher performance with respect to GMV-sample. The network method seems instead more robust, being only slightly affected by the estimation method. The results in Tables 1 show that the network based portfolio selection leads to higher out-of-sample performance with respect to the classical approach.

5 Conclusions

The results obtained clearly show that the network portfolio selection leads to higher out-of-sample performance compared to the classical portfolio allocation. More-

⁸ For space constraint we report only the results obtained in case of the rolling window 2 years in-sample and 1 month out-of-sample

Optimal Portfolio Selection via network theory in banking and insurance sector

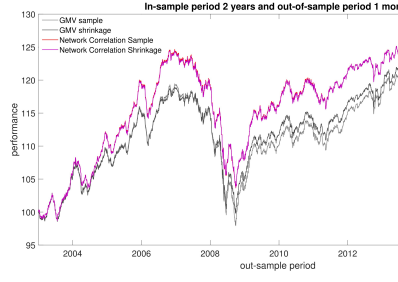


Fig. 2: Out-of-sample performance

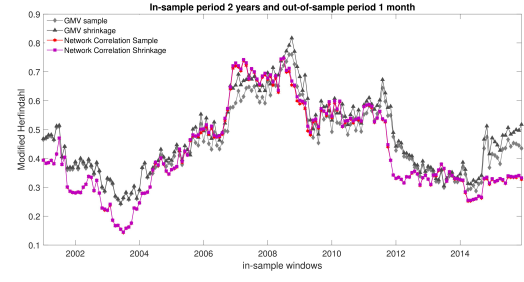


Fig. 3: Modified Herfindahl Index

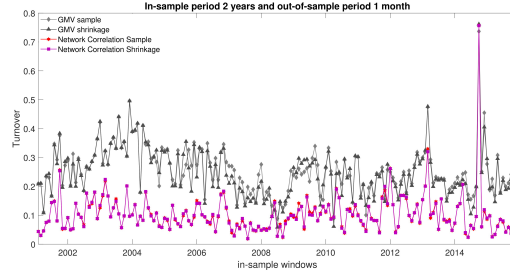


Fig. 4: Portfolio Turnover

Monthly stepped two-years windows	μ_p^*	σ_p^*	skewness	kurtosis	SR	OR	IR
GMV Sample	0.015	0.029	-0.448	9.875	0.525	1.100	
GMV Shrinkage	0.015	0.029	-0.363	9.686	0.533	1.102	0.004
NB Sample	0.015	0.030	-0.517	8.569	0.504	1.093	0.000
NB Shrinkage	0.015	0.030	-0.504	8.525	0.501	1.093	0.000
Monthly stepped six-month windows	μ_p^*	σ_p^*	skewness	kurtosis	OR	IR	
GMV Sample	-0.007	0.032	-0.920	11.774	0.961		
GMV Shrinkage	-0.005	0.030	-0.664	11.860	0.969	0.015	
NB Sample	-0.001	0.033	-0.104	9.939	0.995	0.014	
NB Shrinkage	-0.001	0.033	-0.104	9.939	0.995	0.014	

Table 1: Out-of-sample statistics for Banks & Insurers portfolio with rolling window, respectively, 2 years in-sample and 1 month out-of-sample (upper part of the table) and 6 months in-sample and 1 month out-of-sample (lower part of the table). The statistics μ_p^* and σ_p^* are reported on annual bases. For Monthly stepped six-month windows, values of SR are not reported because of μ_p^* lower than zero.

over, we obtain that the network portfolio selection is robust, in the sense that small changes in the estimated parameters do not lead to drastic changes in the portfolio selection as happen when the classical approach is used.

References

1. Barrat, A., Barthélemy, M., Pastor-Satorras, R., Vespignani, A.: The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences* **101**(11), 3747–3752 (2004)
2. Bloomberg: Bloomberg professional. [online]. Available at: Subscription Service (Accessed: 30 November 2012) (2012)
3. Bongini, P., Clemente, G., Grassi, R.: Interconnectedness, G-SIBs and network dynamics of global banking. *Finance Research Letters* (2018)
4. Cerqueti, R., Ferraro, G., Iovanella, A.: A new measure for community structure through indirect social connections. *Expert Systems with Applications* **114**, 196–209 (2018)
5. Clemente, G., Grassi, R.: Directed clustering in weighted networks: A new perspective. *Chaos, Solitons & Fractals* **107**, 26–38 (2018)
6. Clemente, G., Grassi, R., Hitaj, A.: Asset allocation: new evidence through network approaches. *Annals of Operations Research* (2019)
7. Elton, E., Gruber, M.: Estimating the dependence structure of share prices - implications for portfolio selection. *The Journal of Finance* **28**(5), 1203–1232. (1973)
8. Fagiolo, G.: Clustering in complex directed networks. *Physical Review E* **76**(2) (2007). DOI 10.1103/physreve.76.026107
9. Giudici, P., Spelta, A.: Graphical network models for international financial flows. *Journal of Business & Economic Statistics* **34**(1), 128–138 (2016)
10. Hitaj, A., Zambruno, G.: Are smart beta strategies suitable for hedge fund portfolios? *Review of Financial Economics* **29**, 37–51 (2016)
11. Hitaj, A., Zambruno, G.: Portfolio optimization using modified herfindahl constraint. In: *Handbook of Recent Advances in Commodity and Financial Modeling*, pp. 211–239. Springer (2018)
12. Jobson, J.D., Korkie, B.: Estimation for markowitz efficient portfolios. *Journal of the American Statistical Association* **75**(371), 544–554 (1980)
13. Ledoit, O., Wolf, M.: Honey i shrunk the sample covariance matrix. *Journal of Portfolio Management* **31**(1) (2003)
14. Mantegna, R.N.: Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems* **11**(1), 193–197 (1999)
15. Markowitz, H.: Portfolio selection. *The journal of finance* **7**(1), 77–91 (1952)
16. Martellini, L., Ziemann, V.: Improved estimates of higher-order comoments and implications for portfolio selection. *The Review of Financial Studies* **23**(4), 1467–1502 (2009)
17. McAssey, M.P., Bijma, F.: A clustering coefficient for complete weighted networks. *Network Science* **3**(2), 183–195 (2015)
18. Merton, R.C.: On estimating the expected return on the market: An exploratory investigation. *Journal of financial economics* **8**(4), 323–361 (1980)
19. Minoiu, C., Reyes, J.: A network analysis of global banking: 1978–2010. *Journal of Financial Stability* **9**(2), 168–184 (2013). DOI 10.1209/0295-5075/115/18002
20. Onnela, J., Chakraborti, A., Kaski, K., Kertesz, J., Kanto, A.: Asset trees and asset graphs in financial markets. *Physica Scripta* **2003**(T106), 48 (2003)
21. Tabak, B., Takamib, M., Rochac, J., Cajueirod, D., Souzae, S.: Directed clustering coefficient as a measure of systemic risk in complex banking networks. *Physica A: Statistical Mechanics and its Applications* **394**, 211–216 (2014). DOI 10.1016/j.physa.2013.09.010
22. Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*. Cambridge University Press, New York, NY. (1994)
23. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-worldnetworks. *nature* **393**(6684), 440 (1998)

Matching error(s) and quality of statistical matching in complex surveys

Errori di matching e qualità del matching statistico in indagini complesse

Pier Luigi Conti and Daniela Marella

Abstract The goal of statistical matching, at a “macro” level, is the estimation of a joint distribution having observed only samples from its marginals. In general, the problem is complicate, due to the lack of identifiability of the joint distribution. As a consequence, an intrinsic matching error will affect every matching procedure. A measure of the (maximal) matching error is introduced, in the presence of possible constraints on the joint distribution function, and its estimation on the basis of sample data is dealt with.

Abstract *L'obiettivo del matching statistico, a livello “macro”, consiste nella stima di una distribuzione congiunta sulla base di dati campionari osservati a partire dalle sue marginali. La principale fonte di complicazione, diretta conseguenza del meccanismo osservazionale adottato, è che la distribuzione congiunta in questione non è identificabile. Di conseguenza, una qualunque procedura di matching sarà affetta da un ineliminabile errore di matching. Nel lavoro si introduce una misura del massimo errore di matching, e si studia la sua stima sulla base dei dati campionari osservati.*

Key words: statistical matching, matching error, constraints

1 Introductory aspects. The problem of statistical matching

The increasing availability of inexpensive databases useful for statistical analysis purposes, has, as an important consequence, that the use of data integration tech-

Pier Luigi Conti
Sapienza Università di Roma, P.le A. Moro 5, 00185 Roma, Italy, e-mail: pier-luigi.conti@uniroma1.it

Daniela Marella
Università Roma Tre, Via del Castro Pretorio 20, 00185 Roma, Italy, e-mail: daniela.marella@uniroma3.it

niques is becoming popular. Usually, such databases come either from administrative sources or from institutional sample surveys, or from other sources (as in the case of big data).

In this paradigm, there is a major source of trouble: each database usually contains only *some* of the variates of interest for statistical analysis, and there is no database containing *all* the variables of interest. A major example arises in studying the relationships between income and consumption expenses for households. In Italy reliable information on households income is provided either by the Banca d'Italia Survey on Household and Wealth (SHIW), which is a two-year sample survey, or by the EU-SILC (Statistics on Income and Living Conditions) annual survey, conducted by ISTAT (Italian National Institute of Statistics). Information on consumption expenses are provided by the ISTAT annual sample Household Budget Survey (HBS). Microdata are freely available, and provide information separately on income and consumption expenses. Unfortunately, there is no database containing, for the same sample of households, joint information for both income and consumption expenses. The same is true for many other EU countries.

To be concrete, let \mathcal{U}_N be a finite population of N units, labeled by integers $1, \dots, N$, and let (X, Y_1, Y_2) be three variables of interest; in general, it will be assumed that X is a k -variate character, whilst Y_1, Y_2 are r_1 -variate and r_2 -variate, respectively. The symbols x_i, y_{1i}, y_{2i} denote the values of X, Y_1, Y_2 , respectively, for unit i ($= 1, \dots, N$) of the population.

Given two p -dimensional vectors a, b with components a_j, b_j , the symbol $a \leq b$ will be used to indicate that $a_j \leq b_j$ for every $j = 1, \dots, p$. In the sequel, we will consider a superpopulation model, where x_i, y_{1i}, y_{2i} are realizations of random variables (r.v.s) X_i, Y_{1i}, Y_{2i} that are independent and identically distributed (*i.i.d.*). Furthermore, the (joint) superpopulation distribution function (d.f.) is denoted by

$$H(x, y_1, y_2) = \mathbb{P}(X_i \leq x, Y_{1i} \leq y_1, Y_{2i} \leq y_2), \quad x, y_1, y_2 \in \mathbb{R}. \quad (1)$$

With a little abuse of notation, the case of a nominal X is also included.

In the sequel, we assume that:

- a sample s_1 of n_1 units is selected *via* a (possibly complex) sample design from \mathcal{U}_N , and the values $(x_i, y_{1i}), i \in s_1$, are observed;
- a sample s_2 of n_2 units is selected, independently of s_1 , *via* a (possibly complex) sample design from \mathcal{U}_N , and the values $(x_i, y_{2i}), i \in s_2$, are observed.

The probability that the two samples s_1, s_2 overlap is essentially negligible, so that it can be assumed equal to zero. As a consequence, the variables X, Y_1, Y_2 are not jointly observed.

At a “macro” level, the goal of statistical matching consists in using the sample data

$$\{(x_i, y_{1i}); i \in s_1\}, \quad \{(x_i, y_{2i}); i \in s_2\} \quad (2)$$

to estimate the joint p.d.f. (1) of (X, Y_1, Y_2) . From now on, X will be the variables *common* to the two samples; Y_1 will be the variables *specific* for s_1 ; Y_2 will be the variables *specific* for s_2 .

2 Identifiability in statistical matching

The estimation of the joint p.d.f. of (X, Y_1, Y_2) is equivalent to the estimation of (i) the marginal p.d.f. of X , and (ii) the joint p.d.f. of (Y_1, Y_2) conditionally on X . Part (i) is of course easy. Part (ii) is considerably more difficult, because no joint observations of Y_1 and Y_2 are available.

Let

$$Q(x) = H(x, \infty, \infty), \quad p(x) = Q(x) - Q(x^-) \quad (3)$$

be the marginal d.f. of X and the proportion of population units such that $X = x$, respectively. Define further the conditional p.d.f.s

$$H(y_1, y_2 | x) = H(y_1, y_2, x) / p(x), \quad (4)$$

$$F_1(y_1 | x) = H(y_1, \infty | x), \quad F_2(y_2 | x) = H(\infty, y_2 | x). \quad (5)$$

The only quantity that cannot be estimated from sample data is $H(y_1, y_2 | x)$ in (4).

The knowledge of the p.d.f.s (5) does not generally imply the knowledge of (4). The most important exception occurs when Y_1 and Y_2 are independent conditionally on X : $H(y_1, y_2 | x) = F_1(y_1 | x)F_2(y_2 | x)$. This is the *Conditional Independence Assumption* (CIA, for short).

The most important consequence of CIA is that the statistical model for $H(y_1, y_2 | x)$ is identified by the corresponding marginals $F_1(y_1 | x)$, $F_2(y_2 | x)$. Since the marginals are identifiable by the data observation mechanism, the model for $H(y_1, y_2 | x)$ turns out to be identifiable.

Without CIA, if only the p.d.f.s (5) were known, then one could only say that

$$\max(0, F_1(y_1 | x) + F_2(y_2 | x) - 1) \leq H(y_1, y_2 | x) \leq \min(F_1(y_1 | x), F_2(y_2 | x)). \quad (6)$$

The bounds in (6) are the *Fréchet bounds*. The class of all p.d.f. $H(y_1, y_2 | x)$ satisfying (6) is the *Fréchet class* with marginals $F_1(y_1 | x)$, $F_2(y_2 | x)$. In the sequel, it will be denoted by $\mathcal{F}^x(F_1, F_2)$. If no parametric model for $H(y_1, y_2 | x)$ is assumed, then $\mathcal{F}^x(F_1, F_2)$ is the class of all possible distributions of $(Y_1, Y_2 | x)$.

The absence of CIA makes the statistical model for $H(y_1, y_2 | x)$ *unidentifiable* on the basis of the data observation mechanism. This happens because the identification of $F_1(y_1 | x)$, $F_2(y_2 | x)$ is not enough to identify the whole $H(y_1, y_2 | x)$. The lack of identifiability implies, in its turn, that the traditional inferential paradigm for the estimation of $H(y_1, y_2 | x)$ collapses.

Bounds (6) can be improved by extra-sample information, which is rather frequently available consists in some kind of *constraints*. In the sequel, a few kind of constraints are listed.

1. Constraints involving the dependence structure between Y_1 and Y_2 , given X . An example is the *positive quadrant dependence*, which consists in assuming that

$$\mathbb{P}(Y_1 > y_1, Y_2 > y_2|x) \geq \mathbb{P}(Y_1 > y_1|x)\mathbb{P}(Y_2 > y_2|x)$$

Another example is the *MTP₂* (Multivariate Totally Positive dependence of order 2) between Y_1 and Y_2 given X . If $h(y_1, y_2|x)$ denote the density functions of (Y_1, Y_2) given X , then it is assumed that

$$h(y_1, y_2|x)h(y'_1, y'_2|x) \leq h((y_1, y_2) \wedge (y'_1, y'_2)|x)h((y_1, y_2) \vee (y'_1, y'_2)|x)$$

where $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

2. Constraints involving functionals measuring some characteristics of the dependence between Y_1 and Y_2 (given x). An example is the positive association of Y_1 and Y_2 (given X), *i.e.*

$$\text{Cov}(\psi(Y_1, Y_2), \phi(Y_1, Y_2)|x) \geq 0$$

for every pair non-decreasing, real-valued functions ψ, ϕ . Note that *MTP₂* implies positive association. A simpler example is (in the unidimensional case) the positive covariance between Y_1, Y_2 , given X .

3. Logical constraints, namely restrictions on the support of the variables (Y_1, Y_2) given X . Generally speaking, such constraints can be written in the form

$$a_x \leq f_x(y_1, y_2) \leq b_x \tag{7}$$

given $X = x$, where $f_x(y_1, y_2)$ is an increasing function of y_1 for fixed y_2 and a decreasing function of y_2 for fixed y_1 .

Constraints of the form (7) are considered in [1] from a theoretical point of view. An application to statistical matching of households income and consumption expenses, where constraints are based on average propensity to consumption is in [2].

Constraints 1., 2. are essentially new; they could be especially useful in case of parametric models for the distribution of $(Y_1, Y_2)|X$.

At any rate, using constraints essentially reduces the set of all possible distributions of $(Y_1, Y_2)|X$. If no parametric model is assumed, then the statistical model for the distribution of $(Y_1, Y_2)|X$ becomes a (proper) subset $\mathcal{H}^X(F_1, F_2)$ of the *Fréchet class* $\mathcal{F}^X(F_1, F_2)$. The same holds in case a parametric model for $H(y_1, y_2|x)$ is assumed, although in this case the set of all constrained joint distributions of Y_1, Y_2 given X is considerably “tighter” than the nonparametric case.

3 Matching error in statistical matching

As already said, at a “macro” the problems to solve in statistical matching are essentially two:

- statistical estimation of the joint distribution of (Y_1, Y_2) given X ;
- assessment of the reliability of the estimate of the joint distribution of (Y_1, Y_2) given X .

Now, on the basis of statistical data, it is only possible to estimate $Q(x)$, $F_1(y_1|x)$, $F_2(y_2|x)$. Unless special assumptions are made, one can only say that $H(y_1, y_2|x)$ belongs to the set $\mathcal{H}^X(F_1, F_2)$. In the nonparametric case, this is the set of all joint probability distributions of $(Y_1, Y_2)|X$ having marginals $F_1(y_1|x)$ and $F_2(y_2|x)$ and satisfying the imposed logical constraints.

Each p.d.f. in \mathcal{H}^X is a plausible *matching distribution* for Y_1 and Y_2 , given X . The statistical matching problem essentially consists in choosing a matching distribution, *i.e.* a d.f. in the class $\mathcal{H}^X(F_1, F_2)$ that acts as a surrogate of the actual joint p.d.f.. A *matching procedure* is a procedure to choose a d.f. in the class $\mathcal{H}^X(F_1, F_2)$, but with marginal d.f.s estimated on the basis of sample data.

Suppose that $H^*(y_1, y_2|x) \in \mathcal{H}^X$ is chosen as a matching distribution for (Y_1, Y_2) given X , but that the “true” d.f. is $H(y_1, y_2|x)$. The discrepancy between the chosen $H^*(y_1, y_2|x)$ and the actual $H(y_1, y_2|x)$ is the *matching error*. The matching error is an important measure of the quality of the statistical matching, because because the smaller the matching error, the better the quality of a matching procedure.

The most favourable case, that for instance happens under CIA, occurs when the class $\mathcal{H}^X(F_1, F_2)$ collapses into a *single* d.f., *i.e.* when $H(y_1, y_2|x)$ is identifiable on the basis of sample data. In this case the matching error is null.

Unfortunately, in many cases of practical importance the class \mathcal{H}^X does not reduce to a single d.f., and the matching error *could not* be negligible. In spite of this drawback, the study of the matching error is still of fundamental importance, since a small matching error means that the chosen matching distribution $H^*(y_1, y_2|x)$ is close to the true $H(y_1, y_2|x)$. As a consequence, replacing $H(y_1, y_2|x)$ by $H^*(y_1, y_2|x)$ does not produce a large error.

In the nonparametric case, as a *matching error measure* (conditionally on X) we will use the following:

$$ME_x(H^*, H) = \int_{\mathbb{R}^2} |H^*(y_1, y_2|x) - H(y_1, y_2|x)| d[F_1(y_1|x)F_2(y_2|x)]. \quad (8)$$

With a similar reasoning, as an unconditional measure of matching error we can consider

$$ME(H^*, H) = \int_{\mathbb{R}} ME_x(H^*, H) dQ(x). \quad (9)$$

The matching error, in many cases, can be bounded by some quantity that depends only on the marginal d.f.s F_1, F_2 . In the special case of constraints (7) this is

done in [1]. However, the same idea also applies to positive quadrant dependence and, with more effort, also to MTP_2 (which implies positive quadrant dependence, as already said). To see the point, and to avoid too many technicalities, consider the case of Positive Quadrant Dependence, and assume that both Y_1, Y_2 are univariate. Under Positive Quadrant Dependence, we have

$$H(y_1, y_2|x) \geq F_1(y_1|x)F_2(y_2|x)$$

and hence

$$|H^*(y_1, y_2|x) - H(y_1, y_2|x)| \leq \min(F_1(y_1|x)F_2(y_2|x)) - F_1(y_1|x)F_2(y_2|x) \quad (10)$$

from which it follows that

$$\begin{aligned} ME_x(H^*, H) &\leq \int_{\mathbb{R}^2} (\min(F_1(y_1|x)F_2(y_2|x)) - F_1(y_1|x)F_2(y_2|x)) d[F_1(y_1|x)F_2(y_2|x)] \\ &= \Delta^x(F_1, F_2). \end{aligned} \quad (11)$$

Using similar ideas, it is in general possible to write

$$ME_x(H^*, H) \leq \Delta^x(F_1, F_2). \quad (12)$$

In view of (9), a similar bound also holds unconditionally w.r.t. X .

Although the sample data do not allow to estimate the matching error (8), its upper bound (11) can be estimated, because it only depends on the marginal distributions of Y_1, Y_2 , given X .

In a sense, the quantity $\Delta^x(F_1, F_2)$ in (11) represents a measure of the “size” of the class \mathcal{H}^X . The smaller $\Delta^x(F_1, F_2)$, the smaller the matching error, the closer the matching distribution to the actual distribution of (Y_1, Y_2) given X . Hence, $\Delta^x(F_1, F_2)$ can be considered as a measure to assess how reliable is the use of a matching distribution for $(Y_1, Y_2)|X$ as a surrogate of the actual d.f.. In other words, it is a measure of the quality of matching. Its importance, as already remarked, comes from its most important property: it is *estimable* on the basis of sample data.

4 Estimation of a matching distribution: basic ideas and assumptions

The first goal of statistical matching is to construct a matching distribution for (Y_1, Y_2) given X , *i.e.* in choosing a member of the class $\mathcal{H}^X(F_1, F_2)$.

If no parametric model is adopted, the basic idea is to construct first a matching distribution $H^*(y_1, y_2|x)$ as if $F_1(y_1|x), F_2(y_2|x)$ were known, and then to estimate such a distribution on the basis of sample data.

At a sample level, the unknown d.f.s F_h can be estimated on the basis of samples data *via* their Hájek estimators:

$$\hat{F}_h(y|x) = \sum_{i \in S_h} \frac{1}{\pi_{i,h}} I_{(y_{hi} \leq y)} I_{(x_i = x)} \bigg/ \sum_{i \in S_h} \frac{1}{\pi_{i,h}} I_{(x_i = x)}, \quad h = 1, 2. \quad (13)$$

Under fairly general conditions, an estimator $\hat{H}^*(y_1, y_2|x)$ of $H^*(y_1, y_2|x)$ satisfying the constraints and (uniformly) consistent exists. In [1] the case of constraints (7) is dealt with. The adopted approach is based on a preliminary discretization of data. Then joint distribution for Y_1, Y_2 (given X) satisfying the constraints is first constructed, and then modified by the Iterative Proportional Fitting (IPF) algorithm in order to have marginal d.f.s (13).

The case of positive quadrant dependence between Y_1 and Y_2 given X is more demanding. To simplify computations, a preliminary discretization of data is needed. First of all, it is not hard to see that the class of all joint distributions of Y_1, Y_2 (given X) that are positively quadrant dependent and possess marginals F_1, F_2 is convex. The same hold when the marginal d.f.s F_1, F_2 are replaced by their estimators (13). Next, it is also possible to identify the extreme points of such a convex set. By Krein-Milman theorem, every (estimated) matching distribution can be expressed as a mixture of extreme points. Hence, every (estimated) matching distribution $\hat{H}^*(y_1, y_2|x)$ is a mixture of such extreme points. Note that this also allows to explore the whole (estimated) class $\mathcal{H}^X(\hat{F}_1, \hat{F}_2)$. The consistency of the estimator $\hat{H}^*(y_1, y_2|x)$ is established in next proposition.

Proposition 1. *Assume that X is discrete, with support $\text{Supp}(X) = \{x^1, \dots, x^K\}$. Under general regularity conditions, the following statements hold.*

S1. As N, n_h increase:

$$\frac{1}{Np(x)} \sum_{i \in S_h} \frac{1}{\pi_{i,h}} I_{(x_i = x)} \xrightarrow{P} 1, \quad h = 1, 2, \quad \text{as } N \rightarrow \infty.$$

S2. As N, n_h increase:

$$\sup_y \left| \hat{F}_h(y|x) - F_h(y|x) \right| \xrightarrow{P} 0 \quad \forall x \in \text{Supp}(X).$$

S3. There exists a matching distribution $H^(y_1, y_2|x) \in \mathcal{H}^X(F_1, F_2)$ such that, as N, n_h increase:*

$$\sup_{y_1, y_2} \left| \hat{H}^*(y_1, y_2|x) - H^*(y_1, y_2|x) \right| \xrightarrow{P} 0 \quad \forall x \in \text{Supp}(X).$$

5 Estimation of the maximal matching error

The goal of this section is to provides estimates of both conditional and unconditional maximal matching errors. As already remarked, this is a crucial point in assessing the quality of statistical matching.

To estimate the conditional measure (8), a natural approach consists in estimating first the conditional d.f.s $F_h(y|x)$ s in (5), and then in plugging such estimates in (8). Using the estimators (13), the following estimator of the conditional measure of uncertainty $\Delta^x(F_1, F_2)$ is obtained

$$\hat{\Delta}^x = \Delta^x(\hat{F}_1, \hat{F}_2). \quad (14)$$

The estimate (14) is of interest for two reasons. First of all, it can be constructed on the basis of sample data, so that it solves the problem of assessing the quality of statistical matching. In the second place, it is consistent, as established in next Proposition.

Proposition 2. *Under the same regularity conditions of Proposition 1:*

$$\hat{\Delta}^x \xrightarrow{P} \Delta^x \text{ as } N, n_1, n_2 \rightarrow \infty, \forall x \in \text{Supp}(X).$$

In a similar way, a consistent estimator of the unconditional maximal matching error can be constructed. First of all, as an estimator of $p(x)$ we may consider:

$$\hat{p}(x) = \tau \frac{1}{N} \sum_{i \in S_1} \frac{1}{\pi_{i,1}} I_{(x_i=x)} + (1-\tau) \frac{1}{N} \sum_{i \in S_2} \frac{1}{\pi_{i,2}} I_{(x_i=x)} \quad (15)$$

with $0 < \tau < 1$. The choice of τ may be performed by taking the value that minimizes a convenient approximation of the mean squared error of (15).

Next, as an estimator of $\Delta(F_1, F_2)$ we may consider:

$$\hat{\Delta} = \sum_{x \in \text{Supp}(X)} \hat{\Delta}^x \hat{p}(x). \quad (16)$$

Similarly to Proposition 2), it can be shown that

$$\hat{\Delta} \xrightarrow{P} \Delta \text{ as } N, n_1, n_2 \rightarrow \infty.$$

References

1. Conti P L., Marella D., Scanu M.: Statistical Matching Analysis for Complex Survey Data With Applications. *Journal of the American Statistical Association*, **111**, 1715–1725 (2016)
2. Conti P L., Marella D., Neri A.: Statistical matching and uncertainty analysis in combining household income and expenditure data. *Statistical Methods & Applications*, **26**, 485–505 (2017)
3. Hájek J.: Asymptotic Theory of Rejective Sampling With Varying Probabilities from a Finite Population. *The Annals of Mathematical Statistics*, **35**, 1491–1523 (1964)
4. Hájek J.: Sampling from a finite population, Marcel Dekker, New York (1981)
5. Tillé, Y.: Sampling Algorithms. Springer Verlag, New York (2006)

Hotel search engine architecture based on online reviews' content

Un motore di ricerca per gli hotel basato sulle recensioni online

Claudio Conversano, Maurizio Romano and Francesco Mola

Abstract Based on online reviews available on Booking.com we define an ad-hoc classification model to predict reviews as positive/negative from their content. The log-likelihood ratios attributed to each word included in a review, together with its estimated polarity, are used: a) to estimate the numeric score obtained by a review; b) to decompose the estimated score w.r.t. the different business areas of a hotel; c) to define an alternative customer satisfaction measure. The decomposition of a review content into elementary log-likelihood ratios is also the basis of the design of a novel hotel search engine. The latter is able to efficiently detect the most effective hotel as that matching the user desiderata specified as sentences at the beginning of the searching procedure. The proposed approach is evaluated analysing the reviews provided by tourists who stayed in Sardinian hotels.

Abstract *A partire dalle recensioni presenti su Booking.com si definisce un modello in grado di classificare una recensione in positiva/negativa sulla base del suo contenuto. L'insieme dei rapporti di log-verosimiglianza calcolati per ciascuna parola presente in una recensione e la sua polarità stimata sono utilizzati per stimare il punteggio ottenuto dalla recensione, per scomporre tale punteggio definendo degli score per ogni area di business, e per definire una misura alternativa di customer satisfaction. Gli stessi dati vengono utilizzati per la progettazione e implementazione di un motore di ricerca capace di individuare la struttura più adeguata rispetto ai desiderata del potenziale cliente. L'approccio proposto è valutato analizzando il contenuto delle recensioni relative alle strutture ricettive operanti in Sardegna*

Key words: Big Data, Search Engine, Booking, Sentiment Analysis, Customer Satisfaction, Naïve Bayes Classifier

¹ C. Conversano, M. Romano, F. Mola, Università di Cagliari, Dipartimento di Scienze Economiche e Aziendali; email: {conversa@unica.it, romano.maurizio@unica.it, mola@unica.it}

1 Introduction

Big data has been touted as a new research paradigm that utilizes diverse sources of data and analytical tools to make inferences and predictions about reality (Boyd and Crawford, 2012). Particularly, with increasingly powerful natural language processing and machine learning capabilities, textual contents from the Web provide a huge shared cognitive and cultural context and, thus, have been analyzed in many application domains (Halevy et al., 2009). This phenomenon has characterized the tourism sector also, where product review forums about tourism topics have become commonplace, and an increasing number of websites provide platforms for tourists to publicize their personal evaluations and opinions of products and services. This information is of great interest to both companies and consumers. Companies spend a huge amount of money to find customers' opinions and sentiments, since this information is useful to exploit their marketing-mix in order to affect consumer satisfaction. Individuals are interested in others' opinions when purchasing a product or hiring a service.

From the business viewpoint, online reviews including their peripheral cues, such as user-supplied photos and the reviewer's personal information, are intended as means of persuasive communication in order to build credibility and influence user behavior (Sparks et al., 2013). From an operational point of view, this situation has raised many NLP challenges, commonly referred to as Sentiment Analysis, such as subjectivity detection (Wiebe et al., 1999), polarity recognition (Schmunk et al. 2014) and rating inference (Esuli and Sebastiani, 2006). Focusing on product review classification, various approaches have been proposed during the last decade. Most of them only consider the polarity of the opinions (i.e. negative vs. positive) and rely on machine learning techniques trained over vectors of linguistic feature frequencies.

Within the above-described framework, we have retrieved data about reviews of clients hosted in accommodations, hereafter hotels, located in Sardinia whose information is available on Booking.com. A routine to scrape all the important information concerning a hotel listed on Booking.com has been implemented in Python. Next, to define a benchmarking tool for hotels we have processed reviews' content with natural language to comprehend the items leading towards increasing or decreasing customer satisfaction. The results of data analysis were the basis of the design of a search engine based on clients' reviews, that allows its users to select the most suitable hotel on the basis of a (even small) set of keywords, i.e. a sentence. We have chosen Booking.com as a reference platform as the reviews there available come from customers who effectively stayed in a hotel. Booking.com utilizes a score defined in the [2.5, 10.0] interval and each review is split into two parts: a positive comment and a negative one.

Our research has several purposes. First goal is defining an ad-hoc classifier able to classify a comment as positive or negative from its content. The same classifier allows us to quantify the (positive or negative) impact of a specific word within a review. The second goal is developing a prediction model for the score obtained by a hotel on the basis of the reviews reported on Booking.com. Predicted

Hotel search engine architecture based on online reviews' content

scores constitute a benchmarking tool for a hotel to be evaluated. Last but not least, an addition goal is defining a search engine based on clients' reviews that would allow users to select the most suitable hotel w.r.t. their preferences.

For all these goals to be accomplished, we followed three basic steps: 1) data collection: data were scraped from Booking.com and cleaned in preparation of the statistical analysis; 2) reviews' classification: the reviews' content was processed through an ad-hoc defined Naïve Bayes classifier, hereafter Naïve Bayes*, in order to obtain a predicted score and the polarity for each review based on its specific content; 3) search engine design and implementation: it derives from the results obtained in step 2.

The three above-mentioned steps are described in details in what follows.

2 Data collection

A Python extractor has been implemented that, through web scraping, has retrieved all the useful information publicly available on Booking.com. Retrieved data have next been organized into flat tables. They concern 619 hotels operating in Sardinia. For them, it was possible to scrape 66,237 reviews consisting of 106,800 positive and negative comments. Data about each hotel concerns the hotel information usually available on Booking.com (e.g.: type of accommodation, postcode, city, scores about cleaning, comfort, room, restaurant, etc.). Data about each review concerns its content together with the information about the reviewer and the type of customer (e.g.: business trip, pleasure trip, etc.).

Collected data have been cleaned by removing conjunctions, punctuation, numbers and all the stopwords. Next, words with similar meaning have been merged together in macro-words and all the macro-words composing each comment have been joined in the Bag of Words (BoW). Each element of BoW had its own frequency and these macro-words have been subsequently assigned individually to the following reference categories: "bar", "cleaning", "comfort", "food", "hotel", "position", "price-quality-rate", "room", "services", "sleep-quality", "staff", "Wi-fi" and "other". As an example, the words {"breakfast", "restaurant", "lunch", etc.} all belongs to the "food" category. The definition of these reference categories allowed us to process the data in an aggregate manner, as well as to quantify pros and cons concerning the hotel services associated to each category.

3 Reviews' classification

We have implemented an ad-hoc classifier able to predict as accurately as possible a comment as negative or positive based on the words included in its content. This Naïve Bayes* classifier derives from a modification of the original classifier

having the same name and resulted as the best performing one in terms of generalizability among several of the most commonly used classifiers.

The basic features of Naïve Bayes* applied to reviews' content are as follows.

For a specific review r and for each word w ($w \in BoW$), we consider the log-odds ratio of w , $LOR(w) = \log[P(c_{neg}|w)/P(c_{pos}|w)] \approx pres_w + abs_w$, where c_{pos} (c_{neg}) are the proportions of observed positive (negative) comments whilst $pres_w$ and abs_w are the log-likelihood ratios of the events ($w \in r$) and ($w \notin r$), respectively.

Likewise, for the set of J words included in a comment c , the log-odds ratio of c is defined as: $\sum_{(w_i \in J)} pres_{w_i} + \sum_{(w_i \notin J)} abs_{w_i}$.

We have used cross-validation to estimate a parameter τ such that: c has been classified as "negative" if $LOR(c) > \tau$ or as "positive" if $LOR(c) \leq \tau$. The selected value of τ is that minimizing simultaneously both the Type I and the Type II errors. The same approach has been used for the set of K words composing a review r , thus computing $pres_{w_i}$ and abs_{w_i} for all the words appearing and not appearing in the review, and comparing $LOR(r)$ with the value of τ obtained from the classification of the comments into "positive" and "negative".

To benchmark Naïve Bayes*, we compared its prediction accuracy when classifying comments as positive or negative with that of alternative classifiers, in particular: Logistic Regression, Random Forest, standard Naïve Bayes, Decision Trees and Linear Discriminant Analysis. The Naïve Bayes* classifier performed considerably better than competitors as it provided a Matthews correlation coefficient (Accuracy) of 0.813 (0.9111) versus an average value of 0.327 (0.8081) obtained from the alternatives.

Next, we have used the quantities $\sum_{(w_i \in K)} pres_{w_i}$ and $\sum_{(w_i \notin K)} abs_{w_i}$ computed for a review r to predict the score the reviewer who generated r assigned to a hotel on Booking.com. Again, we learned the same set of alternative classifiers and the best performing one was Random Forest, with $MSE = 0.6704$. We have noticed that considering the log-likelihood ratios of categories originated by merging similar words included in the BoW as inputs for the classifier, instead of the single words ($w \in BoW$), considerably improved the accuracy of the classifier. As an additional predictor, the polarity of a review (positive or negative) estimated in the previous step with the Naïve Bayes* classifier has been considered.

Since Booking.com provides scores for each service offered by a hotel, the same predictive approach has been applied for each service thus obtaining predicted scores arising from reviews' content for each hotel service individually.

4 Search engine design and implementation

The output of the Naïve Bayes* classifier is the basis of a search engine allowing a user to select a suitable hotel according to his desiderata. Let us denote this reference dataset with D . It is composed of the individual terms defining the log-likelihood ratio of words included in BoW, the estimated polarity of the review

Hotel search engine architecture based on online reviews' content

(i.e.: that associated to each specific set of words K), and the scores estimated either for the entire review or for each individual hotel service. Importantly, the log-odds ratios (LOR) have been computed also for a specific hotel h by simply considering the set of H words appearing in all the reviews obtained for h . The empirical distribution of the LOR s obtained for the whole set of hotels allowed us to define a qualitative measure of customer satisfaction for a hotel deriving from a set of five disjoint categories that are: "completely unsatisfactory", "not satisfactory", "potentially good", "good" and "excellent".

The search engine has been designed to be able to provide the most suitable hotel for the user once he has specified a sentence reporting the desired characteristics of the "ideal" accommodation. The matching between the input provided by the user and the output returned by the engine is obtained by searching within \mathbf{D} the words composing the user defined content (input) and returning the information about the hotels that have been better reviewed based on that content.

In the very beginning, the user is required to type in just the name of the destination. Immediately, a list of hotels appears on the screen. The hotels are ordered according to the predicted score obtained from Naïve Bayes*. Interactively, as soon as the user types in additional information (e.g.: a sentence like: "a gourmet restaurant") the list is modified based on the newest information arrived to the engine.

5 Concluding remarks

We have proposed an accurate model to classify online reviews available on Booking.com. Our approach is useful for several reasons. It can support hotels in the identification of their strengths and weaknesses and thus in the identification of the strategic factors leading the management towards service improvement. More extensively, this approach allowed us to identify strengths and weaknesses of the destination where several hotels are located and, for this reason, it can be applied to compare customer satisfaction in different areas. Last but not least, starting from the output of an ad-hoc defined Naïve Bayes* classifier it was possible to accurately predict the score available in Booking.com and to decompose it w.r.t. the different hotel business areas.

The output of the classification model led us towards the design and implementation of a hotels' search engine able to detect the most suitable hotels meeting the desiderata of the user through a matching between user defined sentences and reviewers' content.

The present research has been realized within the research project P.I.A. "Realizzazione di una piattaforma ICT a supporto del settore turistico" (RAS, 2007/2013) financed by Regione Autonoma della Sardegna.

References

1. Boyd D., Crawford K.: Critical questions for Big Data. *Information, Communication & Society*, **15**(5), 662–679 (2012)
2. Esuli A., Sebastiani F.: Determining term subjectivity and term orientation for opinion mining. In: *Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pp. 193–200 (2006)
3. Hawley A., Norvig P., Pereira F.: The unreasonable effectiveness of data. *Intelligent Systems, IEEE* **24**(2), 8–12 (2009)
4. Schmunk S., Höpken W., Fuchs M., Lexhagen M.: Sentiment Analysis: Extracting Decision Relevant Knowledge from UGC. In: Xiang Z., Tussyadiah I. (eds) *Information and Communication Technologies in Tourism*. Springer, Cham (2014)
5. Sparks B.A., Perkins H., Buckley R.: Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behaviour, *Tourism Management*, **39**, 1–9 (2013)
6. Wiebe J.M., Bruce R.F., O'Hara T.P.: Development and use of a gold-standard data set for subjectivity classification. In: *Proc. of ACL'99*, pp. 246–253 (1999)

Economic Crisis and Earnings Management: a Statistical Analysis

Crisi Economica e Gestione degli Utili: un'Analisi Statistica

C. Cusatelli, A.M. D'Uggento, M. Giacalone, F. Grimaldi

Abstract. The financial and real crisis has led to a decline in the confidence towards the financial statements as a tool for representing the actual health status of the companies and it has drawn investors' attention to the financial statement values reliability. This work aims at investigating whether, in the Italian market, the precarious macroeconomic conditions and the consequent difficulties suffered by listed companies have constituted, or not, an incentive to implement earnings management policies manipulations. The large period of time (from 2002 to 2016) allows a mapping of the phenomenon that extends from the period before and after the crisis.

Abstract. *La crisi finanziaria e dell'economia reale ha portato a un calo della fiducia nei confronti del bilancio come strumento in grado di rappresentare il reale stato di salute di una società e ha attirato l'attenzione degli investitori sull'affidabilità dei valori di bilancio. Questo lavoro si propone di indagare se, nel mercato italiano, le precarie condizioni macroeconomiche e le conseguenti difficoltà delle società quotate abbiano costituito o meno un incentivo a implementare politiche di manipolazione degli utili. L'ampio periodo di tempo considerato (dal 2002 al 2016) consente una mappatura del fenomeno che si estende dal periodo pre-crisi a quello post-crisi.*

Key words: Economic Crisis, Earnings Management, Beneish Model

1 Introduction

The numerous cases of business disruptions, involving opportunism and accounting fraud by shareholder, directors and managers, that have occurred in different

¹ C. Cusatelli, University of Bari "Aldo Moro"; email: carlo.cusatelli@uniba.it
A.M. D'Uggento, University of Bari "Aldo Moro"; email: angelamaria.duggento@uniba.it
M. Giacalone, University of Naples "Federico II"; email: massimiliano.giacalone@unina.it
F. Grimaldi, University of Bari "Aldo Moro"; email: francesco.grimaldi1@uniba.it

countries over the past two decades along with institutional and context phenomena and with the rise of the 2008 financial crisis, have refocused the attention of academia, professionals and world policy makers on the disclosure processes used by companies and on to the financial statement values reliability (Grimaldi & Muserra, 2017). The above factors have directed the debate on financial communication processes on two lines of research. The first functional to search for solutions that will improve the economic and financial communication processes and thus the quality of the income (Coffe, 2003). The second is employed in searching for the reasons that induce, administrators and managers, to put in place policies for income manipulation (Ronen and Yaari, 2007).

The theoretical and empirical studies, of prevailing Anglo-Saxon matrix, highlighted, particularly the contribution of the governance mechanisms – internal and external – aimed at influencing the processes for determining and communicating the accounting information, with particular reference to the quality of income, through the identification, quantification and, where possible, the mitigation of the activities earnings management. These studies that detect discrepancies in terms of direction and intensity of the relationship between the corporate governance variables used and the quality of the accounting information, measured by the earnings management, take as their corporate model of reference, the public company.

The empirical studies conducted over the last decade that focused on Italian companies, though still limited in quantitative terms, and definitely not in terms of quality, have analyzed several aspects of the relationship between the variables of corporate governance and the earning quality (e.g., Grimaldi & Muserra, 2017). With reference to the relationship between earnings management and the financial crisis, Graham et al. (2005) show that, when the overall economy is down, CEOs make choices that boost earnings and delay the reversal of these actions until the economy recovers. On the other side, several authors show that companies in financial difficulty are likely to commit less in profit management (Trombetta & Imperatore, 2014) This activity is not without costs and a financial crisis is likely to increase the costs associated to earnings manipulations thereby discouraging managers to engage in it (Zang, 2012).

The international literature has developed a broad modeling that allows to verify the existence of earnings management actions by companies. These studies investigate the informative reliability of quantitative-accounting values by analyzing the potentially distorted use of technical discretion in the preparation of economic-financial documents. For an external person, however, it is difficult to identify which discretionary assessments are truly functional in providing a faithful representation of performance and which have been purposely altered to achieve particularistic goals.

This analysis intends to reveal whether the economic and financial crisis that began in 2007 led, or did not, listed companies in the Italian stock market to implement earnings management policies. The different methods can be distinguished in those that use accounting or statistical tools. Among the latter, the model proposed by M. Daniel Beneish in 1999 boasts a high level of credit in the corporate, academic and institutional world as it allows for the simultaneous evaluation of different aspects of company performance using accounting and extra-accounting data that can be easily found in public budgets (Giunta et al., 2014).

2 The theoretical model

Beneish developed a statistical model that adopts some financial metrics to identify the extent of a manipulation or at least the preconditions that could push companies to engage in this activity, choosing explanatory variables mainly related to:

- the signals on the future prospects that appear in the academic literature, that is, the manipulation of profits is more probable when the future prospects of the companies are insecure;
- the variables based on cash flows and accruals proposed by Healy (1985) and Jones (1991);
- the variables taken from the research of positive theory, which hypothesizes incentives based on contracts for the management of profits (Watts and Zimmerman, 1986).

The result was a model that includes the eight following variables.

1. Days Sales in Receivables Index (DSRI): the ratio between the average collection time in the year and the corresponding measure in the previous year. The average collection time is calculated as the ratio between receivables from customers and sales.

2. Gross Margin Index (GMI): the ratio between the gross margin in the previous year and the gross margin in the current year. To quantify the gross margin, the gross profit rate is calculated as the difference between sales and the costs of goods sold, and is divided by sales.

3. Asset Quality Index (AQI): the ratio between non-current assets (other than property, plant and equipment: PPE) and total assets, compared to the previous year. This index highlights the proportion of total assets for which future benefits are potentially less secure.

4. Sales Growth Index (SGI): the ratio between sales in the year and those in the previous one. In itself an increase in sales does not imply that there is manipulation, but growing companies are considered as more potentially at risk of committing book frauds as capital requirements put pressure on managers to achieve certain goals.

5. Depreciation Index (DEPI): the depreciation rate in the previous year compared to the current one, where those rates are calculated as the ratios between depreciation and the sum between depreciation and fixed assets (PPE).

6. Sales, General and Administrative Expenses Index (SGAI): the ratio between selling, general and administrative expenses and sales in the year, compared to the corresponding measure for the previous year.

7. Leverage Index (LVGI): the ratio between the total debts of a company and the total assets in the year compared to the previous year, and measures the change in debt from one year to another.

8. Total Accruals to Total Assets (TATA): calculated as the difference between current assets and liabilities other than liquidity (working capital) and by subtracting depreciation, compared to total assets. For this index an absolute value is measured and not the change between one year and another in the items taken into consideration. The value of this index can be negative when the operating cash flow exceeds the net profits abundantly, that is the accruals are negative.

Tab. 1 summarizes the maximum likelihood unweighted Probit estimation results (coefficients of the model, their statistical significance and the observed average values of the companies) based on a sample of 50 manipulators and 1708 non-manipulators, explained in the following model with significant log-likelihood ratio test ($\chi^2=129.2$, $p\text{-value}<0.001$) and a descriptive validity with pseudo- $R^2=0.371$:

$$\text{M-Score8} = -4.84 + 0.92 \cdot \text{DSRI} + 0.528 \cdot \text{GMI} + 0.404 \cdot \text{AQI} + 0.892 \cdot \text{SGI} + 0.155 \cdot \text{DEPI} + \\ -0.172 \cdot \text{SGAI} - 0.327 \cdot \text{LVGI} + 4.679 \cdot \text{TATA}$$

Table 1: Coefficients, significance and average values of the eight variables of the M-Score8 model

<i>Variables</i>	<i>Coefficients</i>	<i>T-Statistics</i>	<i>Manipulators</i>	<i>Not Manipulators</i>
Constant	-4.840	-11.01	-	-
DSRI	0.920	6.02	1.465	1.031
GMI	0.528	2.2	1.193	1.014
AQI	0.404	3.20	1.254	1.039
SGI	0.892	5.39	1.607	1.134
DEPI	0.115	0.70	1.077	1.001
SGAI	-0.172	-0.71	1.041	1.054
LVGI	-0.327	-1.22	1.111	1.037
TATA	4.679	3.73	0.031	0.018

Source: Beneish M.D., *The detection of Earnings Manipulation*, in «Financial Analysts Journal», 1999.

The literature contemplates an alternative to the eight-variable model, which takes into account only five variables, leaving aside those considered less significant, adopting the same estimate method as Beneish, formulated as follows:

$$\text{M-Score5} = -6.065 + 0.823 \cdot \text{DSRI} + 0.906 \cdot \text{GMI} + 0.593 \cdot \text{AQI} + 0.717 \cdot \text{SGI} + 0.107 \cdot \text{DEPI}$$

For the company on which the potential manipulation is to be investigated, it is possible to calculate an M-Score value after having replaced, for each model variable, the indexes calculated with the specific balance sheet values of the company in question. The interpretation of the result obtained is the following:

- if $\text{M-Score} > -2.22$ there is a high probability of manipulation;
- if $\text{M-Score} < -2.22$ there is a low probability of manipulation.

This threshold is the most widely used and accepted value in the literature that has adopted this model, but it is not the only one proposed, since, in order to discriminate manipulative companies, it is fundamental to choose the cut-off value (threshold value). This choice depends on the cost and consistency of the probabilistic type I and II errors of the model: lowering the cut-off value reduces type I and increases type II errors and certainly for the investor it is more expensive to consider as not altered a balance sheet that is altered indeed (Beneish 2013).

3 Results and discussion

The reasons behind the practice of manipulating company results are the intention to avoid losses, to reduce and/or accentuate profits, or in any case, not to disappoint analysts' expectations. Most of the scholars of the earnings management

phenomenon are interested in understanding if the extent of this practice is correlated to the varying macro-economic scenarios; in this sense, two research hypotheses are formulated:

- the increase in economic difficulties could lead more and more companies to use a high discretion in drafting the financial statements in order to stabilize the results achieved in the various exercises (Mari et al., 2016);
- in times of crisis there may be a general dissuasion from the implementation of budgetary policies due to the more stringent and precise auditing activity in these periods (Mechelli et Cimini, 2012; Kousenidis, 2013).

Our analysis is carried out on a sample of companies listed on the Milan Stock Exchange during the period between 2001 and 2016 which can be divided in three sub-periods, each consisting of five years: before crisis (from 2002 to 2006), during crisis (from 2007 to 2011), after crisis (from 2012 to 2016). The initial sample included 159 companies covering the main sectors, but the final one deals with 92 companies, because financial and insurance companies, banks, utilities and oil and gas companies have been excluded as they have different and specific regulations.

The M-Score8 model has been adopted and all the indexes, representing the independent variables, have been calculated for each company and each of the 15 years. The main descriptives calculated on the huge amount of data collected and processed for the three sub-reference periods are shown in Tab. 2.

Table 2: Beneish model indexes, mean and median values depending on the phase of the Italian economic cycle

<i>Indexes</i>	<i>Before crisis</i>		<i>During crisis</i>		<i>After crisis</i>	
	Mean	Median	Mean	Median	Mean	Median
DSRI	1.157	0.988	1.028	0.999	1.056	0.975
GMI	1.058	1.017	0.919	0.989	1.048	0.978
AQI	1.263	1.000	1.077	0.999	0.996	1.002
SGI	1.179	1.048	1.094	1.015	1.019	1.022
DEPI	1.192	1.005	0.942	0.976	1.113	0.998
SGAI	1.000	1.000	1.000	1.000	1.000	1.000
LVGI	1.148	0.997	1.041	0.999	1.109	1.013
TATA	-0.130	-0.038	-0.037	-0.033	-0.042	-0.038

It becomes interesting to examine the individual variables to understand which specific parameters determine the risk of manipulation. Starting from DSRI, we note that this value decreases in time of crisis and slightly increases in the following period. It should be highlighted that an increase in this index is associated with a greater probability of manipulation; the calculated values of DSRI are closer to the index average value for non-manipulating companies, therefore, it cannot allow us to understand whether it means a suspected revenue inflation or a change in credit policy. In line with it, we can observe that GMI shows that these indexes are more deteriorated in the pre-crisis and post-crisis periods and this puts pressure on the management and therefore the risk of manipulation raises. In fact, in these two periods the values of GMI are higher than the average value of the index for non-manipulative companies, but they are in any case far from the average value of manipulating companies.

AQI is above the value 1 in the pre-crisis period and in the post-crisis period: the

quality of the activities is particularly affected by the pre-crisis period, where the average value of the index calculated in the sample is slightly higher than that of manipulative companies.

SGI shows average values closer to those of non-manipulative companies in the three periods considered, a value higher than the latter only in the pre-crisis period. This index only measures the evolution of revenues, and a marked increase in the latter can create pressure on managers to continually increase them.

DEPI shows particularly significant average values in the pre-crisis and post-crisis periods: they are above the average value of manipulative companies, this means that managers seem to have leveraged the depreciation rate to inflate profits.

Following Beneish suggestions, the values of SGAI are assumed equal to the neutral value of 1 for the entire period analyzed due to the missing of financial data.

LVGI values are above the unit for the years 2002-2016, while in the pre-crisis period the index reaches 1.148, overcoming the average of the manipulators (equal to 1.111). In the two subsequent periods it shows values closer to the average of the non-manipulators (1.037). The particular contraction of this index in the period of crisis reflects the difficulty in accessing credit for companies, therefore the incentives related to debt contracts, that can lead to the manipulation of profits, are limited.

TATA is always negative in the three periods under; this happens when the operating cash flow exceeds the net profits. The component of the accruals is negative, therefore, the risk that the profits have been forged is relatively low.

On the basis of what has emerged so far, a greater tendency towards earnings management policies in the pre-crisis period is evident. All the variables values are then used to calculate the M-Score, that is the value determining whether the probability of manipulation is high or low. For the analysis, two different model versions have been explored: 1) M-Score8, with eight variables with the coefficients estimated by Beneish (1999) and with a cut-off value of -2.22; 2) M-Score5 (IT), adaptation of the model to the Italian context assuming the cut-off of -4.14. This choice to explore two models depends on the consideration that there is no specific re-adaptation of the Beneish model to companies listed on the Italian Stock Exchange. Hence the hypothesis to compare the outcomes in order to accurately frame the relation between earnings management and the economic crisis in the Italian market. Tab. 3 shows the mean and the median values obtained over the period 2002-2016, grouped by economic cycle.

Table 3: M-Score8 mean and median values, and number of potentially manipulative companies

	Before crisis					During crisis				
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Mean	-2.43	-2.30	-3.82	-2.33	-2.60	-2.35	-2.79	-2.48	-2.52	-2.75
Median	-2.43	-2.63	-3.92	-2.44	-2.54	-2.68	-2.98	-2.66	-2.60	-2.66
Manipulators	35	24	9	26	25	25	10	19	22	14
	After crisis									
	2012	2013	2014	2015	2016					
Mean	-2.82	-2.71	-2.30	-2.63	-2.59					
Median	-2.80	-2.67	-2.57	-2.60	-2.56					
Manipulators	14	15	21	19	20					

By observing the average values of the M-Score8, this value never exceeds the cut-off value of -2.22. To specifically highlight the risk of manipulation, we also report the number of potentially manipulative companies per period, obtained by calculating the M-Score8 value for each company of the sample for each of the fifteen years. Throughout the period under investigation, an overall low presence of companies at risk of manipulation is found, and the largest number is in the pre-crisis period. Following, in the period of intense financial crisis, the number of potentially manipulative enterprises decreases and, finally, in the last years of the post-crisis period there is an increase in companies that have probably adopted earnings management policies.

3.1 The Beneish model adaptation to Italian context

Since the Beneish model (1999) has been developed in the US economic context, it may not have the same "predictive power" if applied to the Italian context, characterized by small and medium-sized companies that draw up their budgets according to the Italian Law, supplemented by the national GAAPs of the OIC.

A model adaptation to the Italian context was proposed by some researchers (Giunta F. et al., 2014), who adopted the same method as Beneish to estimate the indexes coefficients in order to obtain the formula of the Manipulation Score for the Italian context. Since the SGI, SGAI and TATA indexes have not been included as they are not significant, the formula is therefore composed of five variables:

$$\text{M-Score5(IT)} = -6.2273 + 0.488 \cdot \text{DSRI} + 0.1871 \cdot \text{GMI} + 0.2001 \cdot \text{AQI} + 0.2819 \cdot \text{DEPI} + 0.6288 \cdot \text{LVGI}$$

In this application to the Italian context, the threshold value was equal to -4.14 therefore, if M-Score5(IT) calculated for each company is higher, the accounting data may have been altered (with that value the model reduces the errors for false positive at 7.14%). The mean and median values of each score for each of the 15 years are shown in Tab. 4 from which it is clear that the M-score5 (IT) does not exceed the cut-off only in 2003 and 2016.

Table 4: M-Score5(IT) mean and median values, and number of potentially manipulative companies

	Before crisis					During crisis				
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Mean	-4.30	-3.95	-4.35	-4.32	-4.20	-4.35	-4.53	-4.35	-4.62	-4.51
Median	-4.45	-4.47	-4.35	-4.43	-4.46	-4.48	-4.59	-4.45	-4.60	-4.48
Manipulators	19	20	20	10	19	9	5	12	2	7
	After crisis									
	2012	2013	2014	2015	2016					
Mean	-4.40	-4.44	-4.39	-4.50	-4.06					
Median	-4.50	-4.51	-4.46	-4.53	-4.45					
Manipulators	7	5	11	5	16					

Looking at the data in detail, there is a low number of manipulative companies throughout the period under investigation, but the largest number can be found in the

pre-crisis period. Following, in the period of intense crisis and in post-crisis, the number decreases. The Beneish M-Score5 (IT) gives a lower number of potentially manipulative companies than that obtained by using the integral eight-variable version.

4 Concluding remarks

The financial and real crisis has led to a decline in the confidence towards the financial statements as a tool for representing the actual health status of the companies and it has drawn investors' attention to the financial statement values reliability. All the variations of the Beneish model adopted in this paper lead to the same conclusion: there is an overall low presence of companies at risk of manipulation throughout the period under investigation, however the most consistent number is recorded in the pre-crisis period. The greater level of attention aroused by the unfavorable economic situation could have discouraged the Italian companies managers from implementing earnings management policies due to the greater risk of being exposed, given that a more thorough and judicious reading of the financial statements was underway.

References

1. Beneish M.D., The detection of Earnings Manipulation, *Financial Analysts Journal*, (1999).
2. Beneish M.D., Lee C., Nichols D.C., Earnings Manipulation and Expected Returns, «*Financial Analysts Journal*», Volume 69, Numero 2 (2013).
3. Coffee, J., C., (2003). Gatekeeper failure and reform: The challenge of fashioning relevant reforms. *Columbia Law and Economics Working Paper no.237*. Available from: SSRN.com/ abstract=447940.
4. Giunta F., Bini L., Dainelli F., Verifica della base informativa per l'analisi di bilancio: le azioni di manipolazione contabile, *Controllo di Gestione*, Vol.2, Fascicolo 11, pp. 5-17 (2014).
5. Graham, J., Harvey, C., Rajgopal, S., 2005. The economic implications of corporate governance reporting. "J. Account. Public Policy" 40 (1–3), 3–73.
6. Grimaldi, F. Muserra, A.L. (2017). "The Effect of the Ownership Concentration on Earnings Management. Empirical Evidence from the Italian Context". *Corporate Ownership & Control*, 14(3-1), 236-248.
7. Healy, P., The effect of bonus schemes on accounting decisions. *Journal of Accounting and Economics* 7, 85-107 (1985).
8. Jones, J., Earnings management during import relief investigations. *Journal of Accounting Research* 29, 193-228 (1991).
9. Kousenidis D.V., La Ias A., Negakis C.I., The effect of the European debt crisis on earnings quality. *International Review of Financial Analysis*, vol.30, pp. 351-362, (2013).
10. Mari, L.M., Terzani S., Agnello M., Iorio S., Il rischio di manipolazione dei bilanci in tempi di crisi: analisi di un campione di imprese. *Rivista Italiana di Ragioneria e di Economia Aziendale*. Vol.1, (2016).
11. Mechelli A., Cimini R., L'effetto dell'introduzione dei principi IAS/IFRS sull'earnings management nei Paesi dell'Unione Europea. *Rivista Italiana di Ragioneria e di Economia Aziendale*. pp.582-595, (novembre-dicembre 2012).
12. Ronen, J., Yaari V., (2008). *Earnings management: Emerging insights in theory, practice, and research*. Springer, New York.
13. Trombetta, M. Imperatore, C. (2014). "The Dynamic of Financial Crises and its Non-Monotonic Effects on Earnings Quality". *Journal of Accounting and Public Policy*, Vol. 33(3): 205-232.
14. Watts, R., Zimmerman, J., *Positive accounting theory*. Prentice Hall, Englewood Cliffs, NJ (1986).
15. Zang, A.Y., 2012. "Evidence on the trade-off between real activities manipulation and accrual-based earnings management". *Account. Rev.* 87 (2), 675–703.

A Comparison of Nonparametric Bivariate Survival Functions

Confronto tra stimatori non-parametrici della funzione di sopravvivenza bivariata

Hongsheng Dai and Marialuisa Restaino*

Abstract In this contribution we focus on the bivariate survival function and investigate the case when both components are subject to left truncation and right censoring. Particularly, we discuss a number of estimators available in literature and compare their performance by means of a simulation study. An application to a real dataset is also illustrated.

Abstract *Nel presente contributo l'attenzione è rivolta alla stima della funzione bivariata di sopravvivenza, nel caso in cui entrambe le componenti sono censurate e troncate. In particolare, sarà effettuato un confronto tra alcuni stimatori disponibili in letteratura per valutarne la performance attraverso uno studio di simulazione. Una possibile applicazione a un dataset di natura medica è illustrata.*

Key words: Bivariate survival function, Bivariate censoring, Bivariate Truncation

1 Introduction

In survival studies, each subject may experience two types of events and the failure times of the same subject may be highly correlated.

Some examples include times to visual loss on the left and right eyes, times of cancer detection in the left and right breasts, and times to HIV-infection and death due to AIDS. In such studies, either or both failure times may not be observed due to censoring and/or truncation.

Hongsheng Dai

Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom, e-mail: hdaia@essex.ac.uk

Marialuisa Restaino

Department of Economics and Statistics, University of Salerno, Via Giovanni Paolo II 132, 84084 Fisciano (Salerno), Italy e-mail: mlrestaino@unisa.it

* Corresponding author.

The estimation of the bivariate survival function in the presence of censoring and/or truncation is of great importance in applications and is useful in predicting the joint survival experience, and plays crucial roles in estimating the degree of dependence, in model building and testing, and in strengthening marginal analyses.

Nonparametric estimators in presence of only censoring have been proposed by [1, 2, 3, 13, 17, 19].

The case when only one component of the bivariate is subject to truncation has been investigated by [6, 7, 8, 9]. A nonparametric maximum likelihood estimator for the bivariate distribution function where both components are randomly truncated is developed by [12, 18]. The most recent works focused on the nonparametric estimation of the bivariate survival function when both components are truncated and censored (among the others see [4, 5, 14, 15, 16]).

In this contribution, we focus on the nonparametric survival function in presence of bivariate truncation and censoring. In particular, thanks to a simulation study and an application on real data sets, we compare the performance of the most used nonparametric estimators of the bivariate survival function.

The paper is structured as follows. In Section 2 we introduce the estimators proposed by Shen and Yan [16] as generalizations of Dabrowska's (Section 2.1) and Campbell and Földes's (Section 2.2) estimators and the class of estimators proposed by Dai et al. [5] (Section 2.3). In Sections 3-4, the main results of the simulation study and data analysis are illustrated. Finally, some conclusions are given in Section 5.

2 Nonparametric estimators for bivariate survival function

Let (T_1, T_2) , (C_1, C_2) and (L_1, L_2) be the bivariate survival times, right censoring and left truncation, respectively. Since the pair of survival times (T_1, T_2) is subject to right censoring by a pair of censoring times (C_1, C_2) , we only observe $Y_k = \min\{T_k, C_k\}$ and $\delta_k = I[T_k \leq C_k]$ for $k = 1, 2$. Moreover, as the pair (T_1, T_2) is also subject to random left truncation by a pair of truncation times (L_1, L_2) , only subjects with $L_1 \leq T_1$ and $L_2 \leq T_2$ can be observed. The observed data are $(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i}, L_{1i}, L_{2i})$ for $i = 1, \dots, n$. We assume that (T_1, T_2) is independent of the censoring and truncation times, but the censoring and truncation times themselves can be mutually correlated

$$G(t_1, t_2) = P(L_1 \leq t_1 < C_1, L_2 \leq t_2 < C_2).$$

The aim is to estimate the bivariate survival function $S(t_1, t_2) = P(Y_1 > t_1, Y_2 > t_2)$.

When $L_1 = 0, L_2 = 0$, i.e. in absence of truncation, a large number of proposals for estimating the bivariate survival function have been made: among these proposals are Campbell and Földes estimator [1], Dabrowska estimator [2], Prentice-Cai estimator [13].

Shen and Yan [16] proposed a generalization of Dabrowska estimator and Campbell and Földes estimator that account for left truncation and right censoring (Sections 2.1-2.2). In Section 2.3 a class of bivariate survival function proposed by Dai et al in [5] in presence of bivariate truncation and censoring is illustrated.

2.1 Generalization of Dabrowska estimator

Following the paper of Shen and Yan [16], the Dabrowska estimate is given by:

$$\hat{S}(u_1, u_2) = \hat{S}(u_1, 0)\hat{S}(0, u_2) \prod_{y \leq u_2, x \leq u_1} [1 - \hat{L}(dx, dy)], \quad (1)$$

where

$$\hat{L}(dx, dy) = \frac{\hat{\Lambda}_{10}(dx, y-) \hat{\Lambda}_{01}(x-, dy) - \hat{\Lambda}_{11}(dx, dy)}{[1 - \hat{\Lambda}_{10}(dx, y-)][1 - \hat{\Lambda}_{01}(x-, dy)]}.$$

Here the $\hat{\Lambda}$'s are estimates of the bivariate cumulative hazard

$$\begin{aligned} \hat{\Lambda}_{11}(dx, dy) &= \frac{\hat{W}_{11}(dx, dy)}{\hat{R}(x-, y-)}, \\ \hat{\Lambda}_{10}(dx, y-) &= \frac{\hat{W}_{10}(dx, y-)}{\hat{R}(x-, y-)}, \\ \hat{\Lambda}_{01}(x-, dy) &= \frac{\hat{W}_{01}(x-, dy)}{\hat{R}(x-, y-)}, \end{aligned}$$

where

$$\begin{aligned} \hat{W}_{11}(dx, dy) &= n^{-1} \sum_{i=1}^n I[Y_{i1} \leq dx, \delta_{i1} = 1, Y_{i2} \leq dy, \delta_{i2} = 1], \\ \hat{W}_{10}(dx, y-) &= n^{-1} \sum_{i=1}^n I[Y_{i1} \leq dx, \delta_{i1} = 1, L_{i2} \leq y < Y_{i2}], \\ \hat{W}_{01}(x-, dy) &= n^{-1} \sum_{i=1}^n I[Y_{i2} \leq dy, \delta_{i2} = 1, L_{i1} \leq x < Y_{i1}], \\ \hat{R}(x-, y-) &= n^{-1} \sum_{i=1}^n I[L_{i1} \leq x < X_{i1}, L_{i2} \leq y < Y_{i2}]. \end{aligned}$$

The $\hat{S}(u_1, 0)$ and $\hat{S}(0, u_2)$ are the self-consistent estimators of the marginal functions $S_1(t_1)$ and $S_2(t_2)$, respectively. In [16], Shen and Yan proposed a complicated iteration procedure for obtaining consistent estimators of $S_k(u_k)$ ($k = 1, 2$).

2.2 Generalization of Campbell and Földes estimator

Let $S_{2|1}(u_2|u_1) = P(T_2 > u_2 | T_1 > u_1)$.

Using the factorization $S(u_1, u_2) = S_1(u_1)S_{2|1}(u_2|u_1)$, a first-step extension of Campbell and Földes estimator is given as follows:

$$\hat{S}_C^{(1)}(u_1, u_2) = \hat{S}_1^{(0)}(u_1) \hat{S}_{2|1}(u_2|u_1), \quad (2)$$

where

$$\hat{S}_{2|1}(u_2|u_1) = \prod_{y \leq u_2} (1 - \hat{\Lambda}_{01}(u_1, dy)).$$

Next, given $\hat{S}_C^{(1)}(u_1, u_2)$, a first-step estimator of $K(x, y)$ can be obtained by:

$$\hat{K}_C(x, y; \hat{S}_C^{(1)}) = \frac{\hat{R}(x-, y-) \hat{p}(\hat{S}_C^{(1)})}{\hat{S}_C^{(1)}(x-, y-)}. \quad (3)$$

Here, they can obtain a second-step estimator of $\hat{S}_C^{(2)}(u_1, u_2)$ by replacing $\hat{S}_1^{(0)}$ of (2) with the weighted Kaplan-Meier estimator of $\hat{S}_1(u_1; \hat{K}_C^{(1)})$. They continue iterating between (2) and (3) until the estimate converges. Let \hat{K}_C denotes the converged estimator of K and $\hat{S}_C(u_1, u_2) = \hat{S}_1(u_1; \hat{K}_C) \hat{S}_{2|1}(u_2|u_1)$ denote the converged estimator of $S(u_1, u_2)$.

2.3 A class of bivariate survival estimator

In [5], Dai et al developed a new estimator for the joint survival function $S(t_1, t_2)$, considering a transformation for the time points (t_1, t_2) at which the survival function $S(t_1, t_2)$ is to be estimated. For any given arbitrary values (t_1, t_2) , they define a transformation from (t_1, t_2) to (z, α) as

$$t_2 = \zeta(t_1, \alpha), \quad z = \int_0^{t_1} \sqrt{1 + \left[\frac{\partial \zeta(u, \alpha)}{\partial u} \right]^2} du,$$

where $t_2 = \zeta(t_1, \alpha)$ means a function (curve) depending on a parameter α .

They transform the target survival function from $S(t_1, t_2)$ to $S(z; \alpha)$, by the following formula

$$\begin{aligned} S(t_1, t_2) &= P(T_1 > t_1, T_2 > t_2) \\ &= P\left(\int_0^{T_1} \sqrt{1 + \left[\frac{\partial \zeta(u, \alpha)}{\partial u} \right]^2} du > z, \int_0^{T_2} \sqrt{1 + \left[\frac{\partial \zeta^{-1}(u, \alpha)}{\partial u} \right]^2} du > z\right) \\ &= P(Z(\alpha) > z) \equiv S(z; \alpha), \end{aligned}$$

where

$$Z(\alpha) = \min \left\{ \int_0^{T_1} \sqrt{1 + \left[\frac{\partial \zeta(u, \alpha)}{\partial u} \right]^2} du, \int_0^{T_2} \sqrt{1 + \left[\frac{\partial \zeta^{-1}(u, \alpha)}{\partial u} \right]^2} du \right\}.$$

The above transformation implies that they only need to find a consistent estimate for $S(z; \alpha)$, which is the same as $S(t_1, t_2)$.

Thus, the product-limit estimator for $S(z; \alpha)$ is

$$\hat{S}(z; \alpha) = \prod_{s \leq z} \left[1 - \frac{N\{s; \alpha\}}{H_{(n)}(s-; \alpha)} \right], \quad (4)$$

where $N\{s; \alpha\} = N(s; \alpha) - N(s-; \alpha)$, and

$$N(ds; \alpha) = n^{-1} \sum_{i=1}^n I[\tilde{Z}_i(\alpha) \in ds, s > V_i(\alpha), \Delta_i(\alpha) = 1],$$

$$H_{(n)}(s; \alpha) = n^{-1} \sum_{i=1}^n I[\tilde{Z}_i(\alpha) > s \geq V_i(\alpha)],$$

with

$$\begin{aligned} \tilde{Z}_i(\alpha) &= \min\{\tilde{Y}_{1i}, \tilde{Y}_{2i}\}, \\ \Delta_i(\alpha) &= \delta_{1i}I[\tilde{Y}_{1i} \leq \tilde{Y}_{2i}] + \delta_{2i}I[\tilde{Y}_{1i} \geq \tilde{Y}_{2i}] - \min(\delta_{1i}, \delta_{2i})I[\tilde{Y}_{1i} = \tilde{Y}_{2i}], \\ V_i(\alpha) &= \max\{\tilde{L}_{1i}, \tilde{L}_{2i}\}, \end{aligned}$$

$$\tilde{Y}_{1i} = \int_0^{Y_{1i}} \sqrt{1 + \left[\frac{\partial \zeta(u, \alpha)}{\partial u} \right]^2} du, \quad \tilde{Y}_{2i} = \int_0^{Y_{2i}} \sqrt{1 + \left[\frac{\partial \zeta^{-1}(u, \alpha)}{\partial u} \right]^2} du$$

and

$$\tilde{L}_{1i} = \int_0^{L_{1i}} \sqrt{1 + \left[\frac{\partial \zeta(u, \alpha)}{\partial u} \right]^2} du, \quad \tilde{L}_{2i} = \int_0^{L_{2i}} \sqrt{1 + \left[\frac{\partial \zeta^{-1}(u, \alpha)}{\partial u} \right]^2} du.$$

3 A simulation study

A simulation study is conducted to examine the performance of two estimators proposed by Shen and Yan in [16] and the class of estimators developed by Dai et al. in [5]. The scenario described by [16] is recalled.

The (T_1, T_2) 's are i.i.d. bivariate exponential distributed with survival function $S(u_1, u_2) = e^{-(u_1+u_2)-\max(u_1, u_2)}$. The V_2 's are i.i.d. exponential distributed with survival function $S_{V_2}(x) = e^{-\lambda_2 x}$. The V_1 's are defined as $V_1 = B + V_2$, where B 's are

i.i.d. exponential distributed with survival function $S_B(x) = e^{-\lambda_b x}$. The C_1 's are defined as $C_1 = D + V_1 + C_2$, such that $P(V_1 < C_1) = 1$, where D 's and C_2 are both exponential distributed with survival function $S_D(x) = e^{-\lambda_d x}$ and $\bar{Q}(c_2) = e^{-\lambda_q x}$, respectively.

With the choice of $\lambda_b = 10.0$, $\lambda_d = 1.0$, $\lambda_g = (2.0, 7.0)$ and $\lambda_q = (0.5, 1.0)$, we have different percentages of truncation and censoring. The replication is 1,000 times. The sample size is set equal to $n = (100, 200)$.

Tables 1-3 list the probability of truncation (denoted by $q = 1 - p$), the probability of censoring $p_c = 1 - P(\delta_{i1} = 1, \delta_{i1} = 1)$, and the biases and root-mean-squared error (denoted by $\text{RMSE} = \sqrt{\text{mse}}$) of the two estimators by Shen and Yan [16], $\hat{S}_D(u_1, u_2)$ and $\hat{S}_C(u_1, u_2)$, and the estimator by Dai et al. [5] \hat{S}_d , at $S(0.148, 0.06) = 0.7$, $S(0.25, 0.193) = 0.5$ and $S(0.708, 0.193) = 0.2$.

The bias and RMSE of the estimator by Dai et al. [5] (under the transformation $T_2 = aT_1$) and those by Shen and Yan [16] are compared. Looking at the results, it can be noted that the bias of estimator \hat{S}_b is always less than that of \hat{S}_C for $S(0.148, 0.06) = 0.7$, $S(0.25, 0.193) = 0.5$ and $S(0.708, 0.193) = 0.2$, while it is greater than that of \hat{S}_D in most of the cases for $S(0.148, 0.06) = 0.7$ and $S(0.25, 0.193) = 0.5$, while it becomes lower for $S(0.708, 0.193) = 0.2$. Then, in terms of RMSE, \hat{S}_b outperforms \hat{S}_D and \hat{S}_C in most of the cases. In particular, for the estimation of $S(0.148, 0.06) = 0.7$ the RMSE of \hat{S}_b is less than those of other two in most of the cases, while for $S(0.25, 0.193) = 0.5$ and $S(0.708, 0.193) = 0.2$ it is always less.

Table 1 Simulation results for bias and RMSE (general case: $S(0.148, 0.06) = 0.7$).

λ_g	λ_q	q	p_c	n	Bias				RMSE	
					\hat{S}_D	\hat{S}_C	\hat{S}_{new}	\hat{S}_D	\hat{S}_C	\hat{S}_{new}
2.0	0.5	0.67	0.29	100	0.034	0.081	0.030	0.189	0.186	0.216
2.0	0.5	0.67	0.29	200	0.006	0.074	0.024	0.173	0.160	0.175
2.0	1.0	0.67	0.47	100	0.009	0.074	0.029	0.193	0.214	0.219
2.0	1.0	0.67	0.47	200	0.003	0.067	0.024	0.177	0.200	0.175
7.0	0.5	0.42	0.27	100	0.007	0.072	0.025	0.157	0.163	0.170
7.0	0.5	0.42	0.27	200	-0.000	0.054	0.012	0.138	0.129	0.134
7.0	1.0	0.42	0.43	100	-0.022	0.037	0.025	0.179	0.189	0.170
7.0	1.0	0.42	0.43	200	-0.016	0.030	0.012	0.155	0.152	0.134

4 Data analysis

In this section a possible application of the bivariate survival function to AIDS blood transfusion data given by [10, 11] is presented. The estimator considered here is that of Section 2.3.

Table 2 Simulation results for bias and \sqrt{mse} (general case: $S(0.25, 0.193) = 0.5$).

λ_g	λ_q	q	p_c	n	Bias			RMSE		
					\hat{S}_D	\hat{S}_C	\hat{S}_{new}	\hat{S}_D	\hat{S}_C	\hat{S}_{new}
2.0	0.5	0.67	0.29	100	0.002	0.100	0.054	0.179	0.182	0.158
2.0	0.5	0.67	0.29	200	0.008	0.091	0.051	0.143	0.139	0.130
2.0	1.0	0.67	0.47	100	0.011	0.107	0.056	0.192	0.178	0.158
2.0	1.0	0.67	0.47	200	0.008	0.106	0.051	0.174	0.170	0.133
7.0	0.5	0.42	0.27	100	-0.036	0.091	0.049	0.157	0.155	0.127
7.0	0.5	0.42	0.27	200	-0.010	0.059	0.039	0.137	0.129	0.103
7.0	1.0	0.42	0.43	100	-0.026	0.074	0.048	0.141	0.157	0.129
7.0	1.0	0.42	0.43	200	-0.020	0.072	0.040	0.133	0.134	0.104

Table 3 Simulation results for bias and \sqrt{mse} (general case: $S(0.708, 0.193) = 0.2$).

λ_g	λ_q	q	p_c	n	Bias			RMSE		
					\hat{S}_D	\hat{S}_C	\hat{S}_{new}	\hat{S}_D	\hat{S}_C	\hat{S}_{new}
2.0	0.5	0.67	0.29	100	0.000	0.056	0.010	0.150	0.167	0.083
2.0	0.5	0.67	0.29	200	0.008	0.048	0.009	0.142	0.153	0.064
2.0	1.0	0.67	0.47	100	0.023	0.049	0.010	0.173	0.191	0.085
2.0	1.0	0.67	0.47	200	0.017	0.047	0.008	0.162	0.176	0.065
7.0	0.5	0.42	0.27	100	-0.009	0.051	0.010	0.102	0.144	0.068
7.0	0.5	0.42	0.27	200	0.003	0.031	0.006	0.089	0.120	0.055
7.0	1.0	0.42	0.43	100	-0.017	0.042	0.011	0.085	0.118	0.070
7.0	1.0	0.42	0.43	200	-0.012	0.028	0.006	0.055	0.079	0.055

Data were collected by CDC data registry. Adults infected with virus from contaminated blood transfusion in April 1978. Event time is the induction time from HIV infection to AIDS. Infection time is time from blood transfusion to HIV infection. Data left truncated because only subjects who develop AIDS after 1982 are unobserved (as HIV unknown before 1982). Data also right truncated because cases reported after July 1, 1986 are not included in the sample to avoid inconsistent data and bias from reporting delay.

This data frame contains the following columns: induction time, i.e. months between HIV infection and development of AIDS (event time of interest), adult indicator of adult (1=adult, 0=child), infection time, i.e. months from blood transfusion date (Apr 1, 1978) to HIV infection, left truncation time, right truncation time, indicator of event occurrence, which is set to 1 since all subjects experience the event.

The Table 4 shows the estimates of $\hat{S}(t_1, t_2)$ for different values of t_1 and t_2 under the transformation $t_2 = at_1$.

Table 4 AIDS dataset. The $\hat{S}(t_1, t_2)$ at the selected time pairs (t_1, t_2) and their estimated standard error (in parentheses), when the data transformation is $t_2 = at_1$.

	41.43	42.86	44.29	45.71	47.14	48.57	50
21.43	0.3974 (0.0307)	0.3684 (0.0305)	0.3684 (0.0306)	0.3208 (0.0295)	0.3169 (0.0294)	0.3090 (0.0292)	0.2891 (0.0286)
22.86	0.3528 (0.0296)	0.3045 (0.0284)	0.3045 (0.0286)	0.2744 (0.0280)	0.2733 (0.0281)	0.2654 (0.0279)	0.2456 (0.0271)
24.29	0.3379 (0.0292)	0.2897 (0.0279)	0.2859 (0.0278)	0.2488 (0.0265)	0.2488 (0.0267)	0.2488 (0.0272)	0.2337 (0.0267)
25.71	0.3119 (0.0286)	0.2637 (0.0271)	0.2599 (0.0269)	0.2228 (0.0255)	0.2191 (0.0253)	0.2165 (0.0253)	0.2011 (0.0247)
27.14	0.2939 (0.0280)	0.2488 (0.0265)	0.2451 (0.0264)	0.2080 (0.0248)	0.2042 (0.0247)	0.2005 (0.0245)	0.1857 (0.0237)
28.57	0.2482 (0.0265)	0.2105 (0.0249)	0.2080 (0.0248)	0.1708 (0.0230)	0.1671 (0.0228)	0.1634 (0.0225)	0.1485 (0.0217)
30.00	0.2482 (0.0263)	0.2081 (0.0247)	0.2057 (0.0246)	0.1699 (0.0229)	0.1671 (0.0228)	0.1634 (0.0225)	0.1485 (0.0217)

5 Conclusion

In this contribution, the performance of some nonparametric estimators for the bivariate survival function when the bivariate data are subject to truncation and censoring were compared. Thanks to the simulation study it has been underlined that the class of estimators proposed by Dai et al. [5] have a better performance of those by Shen and Yan [16].

References

1. Campbell, G., and Foldes, A.: Large-Sample Properties of Non-parametric Bivariate Estimators With Censored Data. In: Proceedings, International Colloquia on Nonparametric Statistical Inference, Budapest, 1980, Amsterdam: North-Holland pp. 23-28 (1982)
2. Dabrowska, D.M.: Kaplan-Meier estimate on the plane. *Ann. Stat.* **16**, 1475–1489 (1988)
3. Dai, H., Bao, Y.: An inverse probability weighted estimator for the bivariate distribution function under right censoring. *Stat. Probab. Lett.* **79**, 1789–1797 (2009)
4. Dai, H., Fu, B.: A polar coordinate transformation for estimating bivariate survival functions with randomly censored and truncated data. *J. Stat. Plan. Infer.* **142**, 248–262 (2012)
5. Dai, H., Restaino, M., Wang, H.: A class of nonparametric bivariate survival function estimators for randomly censored and truncated data. *J. Nonparamet. Stat.* **28**, 736–751 (2016)
6. Gijbels, I., Gürlér, Ü.: Covariance function of a bivariate distribution function estimator for left truncated and right censored data. Discussion paper no. 9703. Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium (1996)
7. Gijbels, I., Gürlér, Ü.: Covariance function of a bivariate distribution function estimator for left truncated and right censored data. *Stat. Sin.* **8**, 12191232 (1998)
8. Gürlér, Ü.: Bivariate estimation with right truncated data. *J. Amer. Stat. Assoc.* **91**, 1152–1165 (1996)

9. Gürlér, Ü.: Bivariate distribution and hazard functions when a component is randomly truncated. *J. Multivar. Anal.* **60**, 20–47 (1997)
10. Klein, J.P., Moeschberger, M.L.: Survival Analysis Techniques for Censored and truncated data. Springer (1997)
11. Lagakos, S.W., Barraj, L.M., DeGruttola, V.: Nonparametric Analysis of Truncated Survival Data with Application to AIDS. *Biometrika* **75**, 515–523 (1988)
12. Huang, J., Vieland, V.J., Wang, K.: Nonparametric estimation of marginal distributions under bivariate truncation with application to testing for age-of-onset anticipation. *Stat. Sin.* **11**, 1047–1068 (2001)
13. Prentice, R.L., Cai, J.: Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika* **79**, 495–512 (1992)
14. Shen, P.: An inverse-probability-weighted approach to estimation of the bivariate survival function under left-truncation and right-censoring. *J. Stat. Plan. Infer.* **136**, 4365–4384 (2006)
15. Shen, P.: A general semiparametric model for left truncated and right-censored data. *J. Nonparamet. Stat.* **19**, 113–129 (2007)
16. Shen, P., Yan, Y.Y.: Nonparametric estimation of the bivariate survival function with left-truncated and right-censored data. *J. Stat. Plan. Infer.* **138**, 4041–4054 (2008)
17. Tsai, W.-Y., Leurgans, S., Crowley, J.: Nonparametric estimation of a bivariate survival function in the presence of censoring. *Ann. Stat.* **14**, 1351–1365 (1986)
18. van der Laan, M.J.: Nonparametric estimation of the bivariate survival function with truncated data. *J. Multivar. Anal.* **58**, 107–131 (1996)
19. van der Laan, M.J.: Efficient estimation in the bivariate censoring model and repairing NPMLE. *Ann. Stat.* **24**, 596–627 (1996)

Predictive Algorithms in Criminal Justice

Algoritmi predittivi e giustizia penale

Francesco D'Alessandro

Abstract This paper aims at offering an overview of the complex current and foreseeable intertwines between criminal law and developments of artificial intelligence systems. In particular, specific attention will be paid to the risks arising from the application of predictive algorithms in criminal justice.

Abstract *Il presente articolo intende offrire una panoramica di quelle che sono le attuali e le futuribili intricate relazioni tra il diritto penale e lo sviluppo dei sistemi di intelligenza artificiale. Speciale attenzione, in particolare, verrà dedicata ai rischi derivanti dall'utilizzo di algoritmi predittivi nella giustizia penale.*

Key words: criminal law, criminal justice, artificial intelligence, algorithms, algorithmic prediction, predictive justice, predictive policing, robotics, technology.

1 Introduction

From its very first inception in the 1950s – with the seminal paper published by Turing [1] and, especially, the *Dartmouth Summer Research Project* organized by ten pioneering scholars sharing interests in neural nets, automata theory and the study of intelligence – and after several «seasons of hope and despair» [2], the research field of artificial intelligence (AI)² is nowadays living, thanks to the

¹ Francesco D'Alessandro, Full Professor of Business and Corporate Criminal Law, Università Cattolica del Sacro Cuore, francesco.dalessandro@unicatt.it

² For the purposes of this article, the term «*artificial intelligence*» will be used broadly, following the definition adopted by the European Commission's High-Level Group on Artificial Intelligence [3] according to which «Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI system can also be designed to learn to adapt their behaviour by analysing how environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforced learning are specific examples), machine

massive quantity of data available and to the extraordinary developments in computer power over the last years, a new “Golden Age”.

Indeed, the possible applications and potential benefits deriving from the actual and foreseeable developments of these systems are countless [4] and fully capable, as recently acknowledged also by the European Parliament, to «reshape multiple industries» [5]. Likewise tangible, however, are the several concerns coming from multiple voices arising across the globe – from researchers of the most prestigious universities [6, 7, 8, 9], to non-profit organisations [10], from the very same tech giants developing AI systems [11], to national legislators [12] and international policy makers [13] – regarding the possible dangers and threats for humans (or even society as a whole) connected to the conceivable malicious use of AI systems or anyhow the impossibility to control – and sometimes even understand how the systems arrived to – the outcomes of their “decisions”.

In this broad and global debate, criminal law and criminal justice are lately starting gaining a central role on the stage: on the one hand, in fact, several – and, from a theoretical perspective, very complex – queries for criminal law scholars come from the abstract possibility that, in a very near future, a wrongful conduct might be committed by a *machine* without the intervention or interaction of an *individual*, thus critically severing the typical and traditional connection between criminal liability and a human action and/or omission; on the other hand, crime prevention and crime enforcement authorities, as well as the criminal justice systems in several jurisdictions, have already started using AI systems to ground central decisions on the final accountability of defendants – or even to calibrate the justness of the sanctions to be deployed against them – bringing to light new and compound issues to be solved in order to respect all the fundamental rights which are at stake when the criminal process is prompted.

Thus, this paper aims at offering in Section 2 an overview of which are the current and foreseeable intertwines and connections between criminal law and artificial intelligence, to subsequently focus the attention in Section 3, in particular, on the analysis of which are the delicate problems from the application of predictive algorithms in criminal proceedings. Section 4 will then set out few conclusive remarks.

2 Criminal Law and Artificial Intelligence: an overview.

The study of criminal law and artificial intelligence poses a four-fold interrogation.

Firstly, AI systems started – and will increasingly – being used in preventing and enforcing criminal conducts and it is thus important to thoroughly understand whether such tools might over-balance security over privacy (2.1).

reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as integration of all other techniques into cyber-physical systems)».

Secondly, it is very likely that AI systems will be used by humans to commit – intentionally or unintentionally – criminal conducts which are going to be harder to intercept and enforce as well as potentially massive in scale: well-thought actions and policies to prevent such scenarios will be then crucially needed (2.2).

Thirdly, due to the fast advances and developments of AI systems over the last years – and especially to the rapid increase in the number of their applications – it is indeed conceivable that conducts prohibited and punished by criminal statutes will be soon technically committed by machines without the intervention of human beings: this might bring legislators to create new (criminal) accountability regimes, like it already happened with the realm of corporate crime (2.3).

Fourthly, AI programs and tools already entered – and will increasingly do so – in criminal justice system (2.4): as we will analyse also in Section 3 more in detail, this raised several important queries concerning individuals’ fundamental rights at stake – i.e., due process, right to a fair trial, non-discrimination, and so forth – in a criminal trial.

2.1 AI for crime detection, prevention and enforcement

Like in a Hollywood blockbuster movies that we would have labelled as futuristic not so long ago, artificial intelligence is already facilitating – and will increasingly support – private companies in crime detection and public authorities in crime prevention and enforcement.

Within the private sector, for example, sophisticated tools based on machine learning are being used to detect cybercrime¹, money laundering² and even insider trading³. But the current and foreseeable AI applications in the public sector to prevent and enforce criminal conducts are indeed the ones that might bring society to overbalance security over privacy and non-discrimination: in fact, a newly published report by the United Nations and the Interpol [17] informs that several are going to be the uses of AI and robotics in law enforcement, distinguishing between

¹ As noted in a recent article on Bloomberg, in fact, «Microsoft, Alphabet Inc.’s Google, Amazon.com Inc. and various startups are moving away from solely using older “rules-based” technology designed to respond to specific kinds of intrusion and deploying machine-learning algorithms that crunch massive amounts of data on logins, behavior and previous attacks to ferret out and stop hackers» [14].

² As observed by two Authors on the World Economic Forum website «[o]ne of the most compelling use cases for AI is the battle against financial crime. AI has two primary benefits for the banks engaged in this battle: it can increase the effectiveness and efficiency of financial crime investigations, and the institution’s risk management. In addition to helping financial institutions avoid risks by complying more effectively with regulations, it has the potential to slash the costs of the challenge – mainly by reducing false positives in monitoring systems and redirecting the efforts of human experts to other, more productive, areas of suspicious activity» [15].

³ As pointed out, again, on Bloomberg «[i]nsider trading involves illegal profiting from nonpublic information by people buying and selling shares [...]. Now the technology that created artificial intelligence is getting good at detecting stock market fluctuations that can only be explained as abnormal, sophisticated and nefarious. Bloomberg algorithms give market participants help identifying unusual activity in stock, bond, currency and derivatives trading. The automated analysis of derivatives like options can also expose otherwise opaque insider trading activity that was once evident only with the fluctuations of the underlying assets of bonds, commodities, currencies and equities» [16].

*concept stage*¹, *prototype stage*², *evaluation stage*³ and *approved for use stage*⁴ developments.

Numerous are, however, the critics around such systems: in 2017, for example, the Brennan Center for Justice challenged in court the New York Police Department refusal to produce crucial information about its use of predicting policing technologies. As observed, in particular, «[p]redictive policing software typically relies on historic policing data, which can replicate and entrench racially biased policing. Combined with a lack of transparency and oversight, these systems may violate individual constitutional rights and evade community efforts to hold police accountable for their actions» [18]. More generally, it has been sustained that some of these technologies, together with violating privacy rights, might bring to systems of mass surveillance [19].

2.2 *The malicious use of AI*

Along with the risks posed by the so-called Lethal Autonomous Weapons – to date, the European Parliament issued a resolution to ban their development and production [20] and in the U.S. a pledge having the same goal has been signed by more than 200 organizations and 3000 individuals [21] – AI systems are likely to be used, intentionally or unintentionally, to commit wrongdoings.

As illustrated by a detailed report issued in February 2018 by researches and scholars of the Future of Humanity Institute (University of Oxford), the Centre for the Study of Existential Risks (University of Cambridge) and Open AI [22], even limiting the study on the technologies currently available or foreseeable in the next five years, AI systems will alter the landscape of security risks for citizens, organizations and states. In particular, they noted, malicious use of artificial intelligence could threaten *digital security* (e.g. through the activity of criminals who instruct machines to hack data or socially hit victims, at human or superhuman levels of performance), *physical security* (e.g. non-state actors weaponizing consumer drones), and *political security* (e.g. through privacy-eliminating surveillance, profiling, and repression, or through automated and targeted disinformation campaigns).

¹ Among others: (i) AI algorithms to identify stolen vehicles; (ii) analytical tools for video and audio analysis; (iii) machine learning for analysis of seized text-based media to identify potential intelligence; (iv) AI tools for better and more fair criminal investigations.

² Worth mentioning are: (i) augmented-reality driven information extraction to support gathering and processing of online crime reports; (ii) face and soft biometrics to detect suspicious behaviour, identify criminals, and search for persons of interests; (iii) forecasting and preventing political protests and criminal activities; (iv) perimeter patrol robots; (v) use of smart tools to identify child pornography.

³ For example: (i) patrol drones for prisons and borders; (ii) audio and video analysis tools to monitor dangerous prisoners; (iii) surveillance drones; (iv) communications robots; (v) machine learning to analyse voices on telephone calls; (vi) surveillance systems to monitor for and detect criminal behaviour.

⁴ Which are (i) AI bot to identify legally privileged information; and (ii) crime anticipation system to predict temporal and spatial features of crime thereby facilitating law enforcement to optimize its resources and ensure an effective police presence.

Indeed, while such scenarios do not pose any particular issue on the possible accountability of the individuals and/or organizations behind the malicious use of AI, several problems will lie in the detection of such conducts (due to increased anonymity) as well as the possible massive scale of such attacks.

2.3 *Crimes committed by AI systems*

From a criminal law standpoint, on the other hand, complex theoretical questions derive from the abstract possibility that a crime might be committed by a machine without the human intervention.

In dealing with this query, applying the traditional theory of the crime – as suggested by some scholars [23] – do not fully convince, because, as recognized by other authors, «artificial intelligence crime [could] undermine existing liability models, thereby threatening the dissuasive and redressing power of the law»: in fact, «existing liability models may be inadequate to address the future role of AI in criminal activities» [24].

For this reason, a bottom-up approach – i.e., starting from a specific issue to then study and construct the most suitable accountability system on a case-by-case fashion – might be the best way to address the issue: indeed, this seems to be the (correct) road taken by the Council of Europe, which recently started an interesting study on autonomous vehicles [25] in order to find legally sound answers to the questions concerning the possible criminal accountability when a completely autonomous vehicle injures or kills a human¹.

2.4 *Artificial intelligence and criminal justice*

Finally, AI is fully entering in the criminal justice system.

While it is undoubtful that AI will help the judicial actors in reconstructing the scene of a crime as well as in collecting evidence², the application of predictive AI algorithms in the criminal process might raise numerous challenges concerning the fundamental rights of accused persons and defendants.

¹ The concept paper of the Council of Europe rightly points out that «the relatively simple question of *who* is to be held criminally liable for harmful consequences as a result of a machine's autonomous decision-making processes unfortunately does not have a simple answer. For in criminal law it is difficult to deal with "criminal behavior" of non-human beings; if AI takes the place in the driver's seat there will be a responsibility gap». In fact, «the legal framework currently applicable to the development and utilization of autonomous vehicles (or other AI deployment) is based on normative principles developed during the pre-digital era. As a result it is unclear in various situations as to how and when responsibility for harm can be determined» [25].

² For example, in the U.S. the National Justice Institute observed how AI might help in public safety video and image analysis «to match faces, identify weapons and other objects, and detect complex events such as accidents and crimes in progress or after the fact». Again, AI systems can be used in DNA analysis and for gunshot detection [26].

By way of example, in April 2019 an 18-year old sued Apple claiming that its facial-recognition software falsely linked him to a series of thefts from Apple stores, causing him to be arrested for crimes he did not commit [27]. But while waiting the developments of this case, the application of predictive algorithms in criminal trials already brought to the attention of scholars and the general public sensitive questions to be answered.

3 The use of algorithmic prediction in criminal justice.

As recalled by recent scholarship «a death penalty defendant in Pennsylvania state court was denied access to the source code for a forensic software program that generated the critical evidence against him: the program's commercial vendor argued that the code is a trade secret. In a federal court in Texas, the federal government claimed that trade secret interests should shield details about how a cybercrime investigation software operates, even though the information was necessary to determine whether warrantless use of the tool had violated the Fourth Amendment. And in a Wisconsin case, the state supreme court rejected a defendant's claim he had a right to scrutinize alleged trade secrets in an algorithmic risk assessment instrument used to sentence him. The court reasoned that no due process violation had occurred in part because the judge's own access to the secrets was equally limited» [28]. While all three cases raise several queries, let's now focus on the last one, the so-called *Loomis* case.

3.1 *The Loomis case*

After being found guilty of two crimes, Loomis sustained that the trial court's use of an algorithmic risk assessment in sentencing violated his due process rights because the methodology to produce such assessment was not disclosed.

In *State v. Loomis*¹, however, the Wisconsin Supreme Court rejected such a claim; brought to the attention of the U.S. Supreme Court the defendant was denied the petition for a writ of certiorari on June 26, 2017. The assessment was based on a (secret) algorithm called COMPAS, designed by the private company Northpointe to provide decisional support for the Department of Corrections when making placement decisions, managing offenders and planning treatment. The COMPAS risk assessment on the defendant's possible recidivism was based upon information gathered from the defendant's criminal file, and a 137-question interview (some of which, very debatable) with the defendant. Problem being, however, that given the fact that COMPAS was based (also) on historical data, it raised critical issues concerning the racial-based discrimination of its outcomes which have been even recognized by the Wisconsin Supreme Court.

¹ 881 N.W.2d 749 (Wis. 2016)

3.2 *ProPublica's stance, algorithmic (un)fairness and the black box inexplicability*

More specifically, the Wisconsin Supreme Court cited the studies conducted by the non-profit organization ProPublica, which harshly criticized the application of COMPAS in evaluating the risk of recidivism due to several machine biases, also based on the historical data analysed.

They noted, in particular, that «the score proved remarkably unreliable in forecasting violent crime: only 20 percent of the people predicted to commit violent crimes actually went on to do so. When a full range of crimes were taken into account [...] the algorithm was somewhat more accurate than a coin flip» [29]. Additionally, they then turned up (and found) significant racial disparities: «in forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways. The formula was particularly likely to falsely flag black defendants as future criminals» [29].

But as noted by MIT researchers, this problem will probably emerge again due to the fact that «machine-learning algorithms use statistics to find pattern in data. So if you feed it historical crime data, it will pick out the patterns associated with crime» [30]. But the problem, as it is easy to understand, is that those patterns are statistical correlations, nowhere near the same as causations. For example, if an algorithm found that poor schooling was correlated with high recidivism, it would leave you none the wiser about whether poor schooling actually caused crime. So, can we really rely on tools that, as those concerning risk assessment, that act exactly in this way, turning correlative insights into causal scoring mechanisms? It is easy to answer this question negatively, at least for what concerns decisions capable of affecting, in the criminal trial, the freedoms and fundamental rights of individual.

This because the populations that have historically been disproportionately targeted by law enforcement – especially low-income and minority communities – are at risk of being slapped with high recidivism scores. As a result, the algorithm could amplify and perpetuate embedded biases and generate even more bias-tainted data to feed a vicious cycle.

An additional problem, then, comes from the *black box*: i.e., the fact that most of the time – together with being algorithms designed by private companies and protected as trade secrets – «they are given input and produce output, but the exact functional processes that generate these outputs are hard to interpret even to the very scientists who generate the algorithms themselves» [31]. Thus, it might all come down to the concept of *fairness* of the algorithm, which probably raises more issues than the ones that it solves: in fact, «even if a perfectly algorithm does exist, the fairness-as-accuracy definition might still come up short in the event that an algorithm leads to generalization about a particular group» [32].

4 Conclusions.

To conclude, while the benefits deriving from the developments of AI systems are undeniable, likewise concrete are the issues that such systems will bring to the legal system and, especially, to criminal law and criminal justice. It will thus be crucial to properly address such questions, working together with technicians and field experts to find the best solutions: to “steal” an image from Federico Stella [33], we will need to look on how a judoka fights compared to a sumo wrestler, so to use the strengths of technology developments to create a better criminal justice system, instead of blindly trusting machines assessments without proper understanding of what there is behind (or inside) the black box.

References

1. Turing, A.: Computer Machinery and Intelligence. *Mind* (1950), Vol. LIX, Iss. 236, pp. 433-460.
2. Bostrom N.: Superintelligence. Paths, Dangers, Strategies, p. 5. Oxford University Press, Oxford (2014).
3. European Commission’s High-Level Expert Group on Artificial Intelligence: A Definition of AI. Main Capabilities and Scientific Disciplines (18 December 2018), Brussels.
4. Stone P. *et al.*: Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence (September 2016), Stanford.
5. European Parliament: Resolution on a comprehensive European industrial policy on artificial intelligence and robotics (12 February 2019), Strasbourg.
6. Russell S. *et al.*: Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine* (2015), Vol. 36, Iss. 4, pp. 105-114.
7. Barrat J.: Our Final Invention. Artificial Intelligence and the End of the Human Era. Thomas Dunne Books, New York (2013).
8. Bostrom N. *et al.*: Unprecedented Technological Risks (September 2014), Oxford.
9. Stoica I. *et al.*: A Berkeley View of System Challenges for AI (2017), Berkeley (CA).
10. World Economic Forum: Harnessing Artificial Intelligence for the Earth (January 2018), Davos.
11. Pichai S.: AI at Google. Our Principles (7 June 2018), Mountain View (CA).
12. Agenzia per l’Italia Digitale: L’intelligenza artificiale al servizio del cittadino (10 March 2018), Rome.
13. United Nations: United Nations Activities on Artificial Intelligence (2018), New York.
14. Bass D.: Artificial Intelligence vs. the Hackers (3 January 2019) Bloomberg.
15. Zimiles E., Mueller T.: How AI is transforming the fight against money laundering (17 January 2019), World Economic Forum.
16. Winkler M.: To Sniff Out Insider Trading, Follow the Options Market (28 November 2018), Bloomberg.
17. United Nations, Interpol: Artificial Intelligence and Robotics for Law Enforcement (2019), Turin – Lyon.
18. Levinson-Waldman R., Posey E.: Predictive Policing Goes to Court (5 September 2017), Brennan Center for Justice.
19. Human Rights Watch: China: Police ‘Big Data’ Systems Violate Privacy, Target Dissent (19 November 2017), New York.
20. European Parliament: Resolution on autonomous weapon systems (12 September 2018), Strasbourg.
21. Future of Life Institute: The Risks Posed Lethal Autonomous Weapons Pledge (4 September 2018) Cambridge (MA).
22. Brundage M. *et al.*: Malicious Use of Artificial Intelligence. Forecasting, Prevention, Mitigation (February 2018), Oxford.
23. Hallevy G.: Liability for Crimes Involving Artificial Intelligence Systems (2015), Springer, Berlin.
24. King T. *et al.*: Artificial Intelligence Crime. An Interdisciplinary Analysis of Foreseeable Threats and Solution (2018), Oxford.

Predictive Algorithms in Criminal Justice

25. Council of Europe: Artificial Intelligence and Criminal Law Responsibility in Council of Europe Member States. The Case of Autonomous Vehicle (14 September 2018), Strasbourg.
26. Rigano C.: Using Artificial Intelligence to Address Criminal Justice Needs (8 October 2018), National Institute of Justice.
27. Van Voris B.: Apple Face-Recognition Blamed by N.Y. Teen for False Arrest (23 April 2019), Bloomberg.
28. Wexler R.: Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System. *Stanford Law Review* (May 2018), Vol. 70, pp. 1343-1429.
29. Agwin J. *et al.*: Machine Bias (23 May 2016), ProPublica.
30. Hao K.: AI is sending people to jail – and getting it wrong (21 January 2019), MIT Technology Review.
31. Rahwan *et al.*: Machine behaviour (25 April 2019), *Nature*, Vol. 568, pp. 477-486.
32. Kehl D., Kessler S.: Algorithms in Criminal Justice System. Assessing the Use of Risk Assessment in Sentencing (2017), Harvard Law School.
33. Stella F.: Criminalità d'impresa: lotta di sumo e lotta di judo, *Rivista Trimestrale di Diritto Penale dell'Economia* (1998), vol. XI, pp. 459-477).

A proposal for an integrated approach between sentiment analysis and social network analysis

Una proposta per un approccio integrato tra analisi del sentimento e analisi delle reti sociali

Domenico De Stefano and Francesco Santelli

Abstract Classical approach in opinion diffusion studies are usually based on performing separately social networks or textual sentiment analysis. Few contributions are focused on a simultaneous approach that exploits the interactions between these two methods. The present work proposes an approach to combine social network analysis and sentiment analysis tools in order to explore the determinants of opinions diffusion in social online communities, namely on Twitter. We use a multi-steps procedure to derive signed networks related to the spread of contents and opinions. A comparative analysis will be possible as well among the several concept networks.

Abstract *Gli approcci classici per l'analisi della diffusione delle opinioni si basano spesso sull'utilizzo ben separato di tecniche di analisi delle reti sociali e dell'analisi del sentiment. Pochi contributi si concentrano sull'uso simultaneo e complementare di queste due tecniche. Nel presente lavoro si propone un approccio congiunto tra sentiment analysis e analisi delle reti sociali al fine di studiare la dinamica della diffusione delle opinioni in una particolare comunità online (Twitter). In particolare si tratta di una procedura a step multipli che consente di derivare reti segnate in grado di descrivere la struttura di diffusione di contenuti e opinioni. L'approccio forisce la possibilità di condurre analisi comparative tra differenti strutture che caratterizzano i diversi concetti.*

Key words: Social Networks, Sentiment Analysis, Opinion diffusion, Signed networks, Twitter

Domenico De Stefano

Dept. of Political and Social Science, University of Trieste, e-mail: ddestefano@units.it

Francesco Santelli

Dept. of Political and Social Science, University of Trieste, e-mail: fsantelli@units.it

1 Introduction

In Online Social Media Data (OSMD) a key role is played by Twitter platform, especially for what concerns the analysis of opinions (5), opinions in themselves, but more interestingly also the way such opinions spread in the chain of the users of the Twitter network. In this framework, retweets analysis arises as a crucial phase in understanding the mechanism and the dynamics of such opinion flow (6; 8). Retweeting is, indeed, likely the most adopted proxy to investigate how topics move from one user to another. It depends, in first place, from the level of how “influencing” is the user that has written the original tweet. Opinions shared by users with a very high number of followers are likely to be spread in a large retweeting network. However, also other kind of interactions, likewise mentioning a user or reply to a tweet may convey information or opinions on a particular topic. A reply is made when a user answers to the original tweet, leaving his own personal opinion as a comment. This behaviour, and the following interpretation of the replies structure, is strictly linked to the topic of reference. Sometimes users reply in order to express their disagreement with the original meaning of the tweet. But they can also reply to reinforce the tweet concept/topic. Replies that can be considered as fully neutral are not so common, yet still conceivable. The amount of different possible aims behind replies is an element that increases notably the complexity of the opinions analysis, due to this high degree of heterogeneity in Twitter actions behaviour.

In the present paper, we combine social network analysis and sentiment analysis tools in order to explore the determinants of opinions diffusion in the Twitter platform. In particular, we adopt a multi-steps procedure to derive and analyze signed networks, that are structures related to the spread of contents and opinions. Such networks come from two main actions: retweet and reply, and make use of sentiment analysis to determine the sign of each link. The general idea is first of all to reduce the problem dimensionality, with a clustering procedure on tweets based on the concepts they express. Each group of tweets will spread the related concepts by means of the previous mentioned interactions (retweets and replies) in a way determined by the configuration of the associated signed network.

The contribution is organized as follows: in section 2 we motivate the use of both retweet and reply interactions for studying the opinion flow; in section 3 we describe the proposed procedure that allows to reconstruct the opinion flow signed network; section 4 contains preliminary results of a case study; section 5 concludes.

2 Replies and retweets interactions to describe opinionflow

In this work, we reconstruct the tweet-retweet and tweet-reply relations of opinions about a trending topic on the Twitter platform implementing some elements related to the semantic field of tweets (7). The reason why we adopt both these forms of interaction is that while for retweeting the content is usually fixed and no further content is added, for what concerns replies there is a particular kind of conditional

semantic relationship between the reply and the original concept expressed in the replied tweet. In analyzing the spreading of opinions, we adapt different sentiment analysis algorithms (1) in order to determine the sign of both the original tweet (with respect to the studied topic) and the sign of the edge connecting the original tweet to the replies.

In our framework, the sentiment analysis leads to a signed network of tweet-retweet and tweet-reply relationships (9). For Twitter data, several procedures and algorithms in order to understand the sentiment of tweets have been developed (4) however, we assume that the sentiment attached to each reply depends not only upon the trending topic but also upon the replied tweet. The aim is, on the other hand, to find a semantic pattern in the retweeting mechanism. The insight is indeed to evaluate a potential relationship between such semantic field, broadly defined, and the amount of retweeting/replying of the original tweet, considered as the proxy of the power of the tweet outgoing opinionflow.

In the following we describe the multi-step procedure to derive and analyze the associated signed networks of opinion spreading.

3 Constructing opinion flow signed network

3.1 *Tweets dimensionality reduction step*

The approach to reduce the dimensionality and complexity of original tweets corpus is composed of several phases. Starting from the Twitter data, a network of hashtags is extracted, in a symmetric adjacency matrix $\text{hashtags} \times \text{hashtags}$. Such network is weighted (number of times that hashtags occur in the same tweet or retweet), undirected and, as mentioned, symmetric. In this proposed network, a community detection algorithm (fastgreedy) is performed (3) and clusters of hashtags highly connected are detected. Each group of hashtags defines a concept, or at the very least, a trending topic about the original key hashtag selected to collect tweets. Tweets and retweets are then marked by the concepts they include inside them. Some tweets will include only one concept, but some of them will include several of them. A hierarchical clustering procedure is therefore applied to tweets once properly described by concepts (2). In this phase, the step is made to reduce dimensionality of the original corpus: from a very large number N of tweets to a k number of clusters of tweets described by concepts, with $k \ll N$. Then, the network of tweets-retweets is drawn. Original tweets will belong to different clusters and each cluster will be denoted, as said, by different concepts. It will be shown that some groups of tweets, characterized by some peculiar concepts with a strong polarization (similarly to a sentiment analysis view), will be able to encourage the retweeting mechanism more than others groups, denoted instead by more neutral, or less extreme, concepts.

3.2 *The role of sentiment analysis*

Next step consists in determine, starting from the detected semantic field of concepts attached to the original tweets, the signed network derived from the (conditional) sentiment embedded in both retweet and replies. The tweet-retweet relations will spread the same concepts embedded in the original tweet (that can be positive, neutral or negative with respect to the considered topic). Whereas, in the case of replies, the sign (sentiment) will not necessarily be the same of the original concept of the replied tweet. It is in fact pretty common to express a strong disagreement within a reply. Therefore, the resulting signed network will include both retweeting and replies: each retweet will keep the original sign of the tweet, while the conditional sentiment will determine the sign of the reply, that will be basically agreement (+), disagreement (-) or neutral attitude. We will obtain a number of signed networks for each tweet conveying a given concept or group of concepts. A further step will consist on the classification of the different signed network structures obtained in the previous step. Each network embeds the information and the structure of how positive, negative or neutral concepts spread. It will be shown how these derived network structures act and how they are crucial in understanding opinion spreading, taking under consideration the relationships between signs, concepts and network configurations.

4 Case study: tweets related to the hashtag *flattax*

To give a preliminary illustration of how the proposed procedure works, it will be presented an application on real data. It will be deepen a topic that is widely discussed in Italy, that is the *flat tax* issue. A flat tax system applies the same tax rate to every citizen regardless of income bracket. Italian Government aims to establish this tax system, mainly under the pressure of Lega Nord (Northern League) and its political leader Matteo Salvini. This has been indeed one of the most used leitmotif of such party during last electoral campaign. The system is of course not easily applicable due to its economic cost for public expenditure, and that is on the reason why there is a strong public debate around it. In the last Economic and Financial Document of the 9th April, 2019, it has been somehow introduced officially in the Italian system, even if not in the fully extent envisaged by Matteo Salvini. For this, the majority of newspapers have written that this flat tax is now in its *embryonic stage*, and will be gradually and slowly introduced in the system in the future.

Data related to this issue are collected using as key the hashtag #flattax: only tweets denoted by this hashtag are included in the analysis. The temporal window goes from the early morning of 9th April, 2019, to the afternoon of 21st April, 2019. Tweets including hashtag #flattax in this range are collected by means of the free Twitter API. The total number of tweets is 8258, including retweets and replies. Replies are 304 and retweets are 6672.

It is worth to point out that words sharing same colors in fig.1 describe also a peculiar political affiliation: for example, all the light-blue hashtags are related to politicians belonging to Democratic Party. It is, of course, denoted by tweets in strong disagreement with respect to flat tax. On the other hand, purple words are expression of support to the government action: #governodelcambiamento (government of change) is one of the most used catch-phrase by members of Lega Nord and Movimento 5 Stelle (Five Star Movement), the two main parties of the actual government.

We will use these two concepts as benchmarks of both positive and negative attitude with respect to flat tax. Tweets containing mostly light-blue hashtags are 285: 15 original tweets, 268 retweets and 2 replies. Tweets containing mostly purple hashtags are 940: 55 original tweets, 872 retweets and 13 replies. In both cases, retweets play the role of the most important tool to agree and reinforce the concept expressed in the original tweet. The average number of retweets for light-blue group is 88, while for purple group is 124. However, shrinking analysis only to original tweets, the most retweeted are the ones belonging to light-blue group (21 on the average), while the mean number of retweets for tweets belonging to purple group is only 13.

Replies are very rare in this network, and for the most part they stress out a disagreement (-), giving a negative sign to the link. For example, a reply to the original tweet of Matteo Salvini giving his endorsement for the flat tax institutions, a reply that has been added to favorite 15 times, says: “Sei un evasore? Tranquillo, ci sar la #pacefiscale” (Are you a tax-evader? Keep calm, there will be the #fiscalfare). That means that the use of sarcasm is very popular on twitter when replying, adding elements to the sentiment analysis phase to be investigated.

5 Further developments

In this work we have shown how to use hashtags relationships, in terms of social network analysis, to find out communities of hashtags related to latent semantic concepts. Further is has been possible to use such concepts to cluster tweets, and understand how different groups of tweets behave in terms of retweets and replies: that is, in term of spreading the agreement or disagreement about the embedded concepts. In the future we will increase the number of collected tweets in order to obtain a much greater number of replies, leading to a way complex corpus to analyze using appropriate sentiment analysis algorithms. From that, it will be possible to construct different signed networks for each concepts (or even for individual tweets), making available a exhaustive analysis of concepts diffusion on Twitter.

References

- [1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 30–38 (2011).
- [2] Balbi, S., Misuraca, M., Spano, M.: A two-step strategy for improving categorisation of short texts. Proceedings of JADT conference (JADT 2018), pp. 60–67 (2018).
- [3] Fortunato, S.: Community detection in graphs. *Physics reports* **486**, 75–174 (2010)
- [4] Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision, In: CS224N Project Report, Stanford, **1**(12) (2009).
- [5] Onorati, T., Díaz, P.: Giving meaning to tweets in emergency situations: a semantic approach for filtering and visualizing social data, *SpringerPlus* **5**(1), 1782 (2016).
- [6] Rossi, L., Magnani, M.: Conversation practices and network structure in Twitter. In: Sixth International AAAI Conference on Weblogs and Social Media (2012).
- [7] Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: International semantic web conference, 508–524. Springer, Berlin, Heidelberg (2012).
- [8] Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: 2010 IEEE Second International Conference on Social Computing, 177–184 (2010).
- [9] Tang, J., Chang, Y., Aggarwal, C., Liu, H.: A survey of signed network mining in social media, *ACM Computing Surveys (CSUR)* **49**, 42 (2016)

A meta-tissue non-parametric factor analysis model for gene co-expression

Meta-analisi fattoriale non parametrica per lo studio di espressioni genetiche in diversi tessuti

Roberta De Vito and Barbara Engelhardt

Abstract Tissue-specific patterns are becoming crucial for understanding molecular mechanism of tissue heterogeneity. In this framework, common genes pattern express in all tissues can mask the signal specific to the tissue. We address these challenges in the context of extracting co-expression gene clusters essential to the identification of patterns across tissues coming from correlated brain regions of overlapping sets of individuals. Our method is a meta-study Bayesian nonparametric generalized latent factor model with three important features a) tissue-specific latent features, b) common latent features, and c) a error term accounting for overlapping subjects. We illustrate the advantages of the proposed method through application to a high dimensional biological data set.

Abstract *I patterns di geni associati a diversi tessuti sono rilevanti per la comprensione di meccanismi molecolari dei tessuti stessi. In questo contesto, componenti di geni comuni a tutti i tessuti possono influenzare e mascherare il vero segnale associato ai tessuti specifici. L'obiettivo di questo manoscritto consiste nell'estrarre cluster di geni essenziali nell'identificazioni di componenti associati a tessuti derivanti da regioni del cervello. Il nostro metodo è uno studio bayesiano non parametrico di fattori latenti con tre componenti a) variabili latenti associate ad ogni tessuto, b) variabili latenti comuni a tutti i tessuti e c) un termine di errore relativo ad ogni soggetto. I vantaggi del metodo sono illustrati attraverso un'applicazione a dati biologici.*

Key words: Big-data, Factor analysis, Tissue

Roberta De Vito
Princeton University, Princeton, NJ, USA, e-mail: rvito@princeton.edu

Barbara Engelhardt
Princeton University, Princeton, NJ, USA, e-mail: bee@princeton.edu

1 Introduction

All cells in the animal or human body have the same DNA, and consequently, cells in different tissues share common genes. However, cells in different tissues show sets of genes that display variations associated with specific tissue [6]. In fact, genes in the human genome can display temporally and spatially variations in gene expressions implying tissue heterogeneity. Although common biological processes are achieved in all human tissues, tissues are differentiated by gene expression patterns, resulting in distinct regulatory pathways controlling tissue specificity [5].

Broad analyses of human gene expression levels can be assessed and detect target selection. Indeed, previous studies showed that genes involved in tissue “signature networks” are twice as likely to become drug targets compared to genes expressed everywhere [10]. Consequently, researches focus on the estimation of tissue-specific gene expression patterns promise to be a rich source of targets and biomarkers for disease treatment and diagnosis [9].

In this picture, it is crucial to understand the tissue-specific biological processes. Therefore, an accurate comprehension of regulation and expression of genes in different tissues can help elucidate the molecular mechanisms of tissue development and function.

Two important considerations need to be done. Firstly, oversimplified models could miss important information regarding the tissue-specific pathway. Specifically, we need to remove the common noise that lacks tissue-specific heterogeneity. Second, we rigorously need to validate the estimated tissue pathways.

In this article, we adopt the Bayesian Multi-study Factor model (BMSFA) [3], to estimate both the tissue-specific factors and the common factors. The BMSFA is a dimension-reduction approach that allows for joint analysis of multiple studies, achieving the goal of capturing common factors and the specific factors. This paper aims to estimate the study-specific signal and systematically remove the common signal. With regard to interpretation, the common signal is more likely to cover the tissue heterogeneity resulting in tissue “signature networks” obscured by common pathway shared among all the tissues.

Particularly, we applied the BMSFA in the Genotype-Tissue Expression (GTEx) project [7]. The GTEx consortium considers a wide variety of human tissues. It includes genotype, gene expression, histological and clinical data for 449 human donors across 44 tissues. This project is crucial to study tissue-specific gene expression and to identify tissue “signature networks” across many diverse tissues.

This paper is the first attempt to extrapolate tissue-specific network and systematically remove the common that can cover the tissue-specific signal.

The plan of the paper is as follows. Section 2 introduces the BMSFA and describes the prior. Section 3 illustrates the application to the GTEx project focusing on brain regions.

2 Methods

In this section, we provide details for the model definition and the prior setting. For clarity, we present the method focus on detecting tissue heterogeneity and simultaneously removing the common component that can impact the tissue heterogeneity. We emphasize that the model can be applied to a variety of other situations where both similarities and differences exist in multiple high dimensional data sets.

We adopt the Bayesian Multi-study Factor model (BMSFA)[3] to study and detect tissue heterogeneity. We consider S studies, each with the same P genomic variables.

Specifically, considering S studies, we assume that

$$\mathbf{x}_{is} = \Phi \mathbf{f}_{is} + \Lambda_s \mathbf{l}_{is} + \mathbf{e}_{is}. \quad (1)$$

with the two common latent vectors $\mathbf{f}_{is} \in \mathbb{R}^{k \times 1}$, $i = 1, \dots, n_s$; the study-specific latent vector for study one $\mathbf{l}_{is} \in \mathbb{R}^{j_s \times 1}$, $i = 1, \dots, n_s$. The common factor loadings $\Phi \in \mathbb{R}^{p \times k}$ captures the shared influence that the common latent vector (\mathbf{f}_{is} , $i = 1, \dots, n_s$) have on all the (\mathbf{x}_{is} , $i = 1, \dots, n_s$, $s = 1, \dots, S$). S additional matrices, $\Lambda_s \in \mathbb{R}^{p \times j_s}$, capture the specific influence that the study-specific latent vectors (\mathbf{l}_{is} , $i = 1, \dots, n_1$) have on the observed variables. Finally, a gaussian noise \mathbf{e}_{is} , $p \times 1$ vector is assumed to have a covariance $\Psi_s = \text{diag}(\psi_{s_1}^2, \dots, \psi_{s_p}^2)$. As a result, the marginal distribution of the observed variables, \mathbf{x}_{is} , is a multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\Sigma_s = \Phi \Phi^\top + \Lambda_s \Lambda_s^\top + \Psi_s$. The BMSFA is not identifiable, and for any orthogonal matrix \mathbf{Q} and \mathbf{Q}_s with $\mathbf{Q} \mathbf{Q}^\top = \mathbf{I}$ and $\mathbf{Q}_s \mathbf{Q}_s^\top = \mathbf{I}_s$ we can obtain the same identical solution for (1) with $\Phi^* = \Phi \mathbf{Q}^\top$, $\Lambda_s = \Lambda_s \mathbf{Q}_s^\top$, and $\mathbf{f}_{is}^* = \mathbf{Q} \mathbf{f}_{is}$, $\mathbf{l}_{is}^* = \mathbf{Q}_s \mathbf{l}_{is}$. We solved the identifiability issue providing different solutions [4, 3]. Here, we focused on the estimation of the common part of the covariance matrix, i.e. $\Sigma_\Phi = \Phi \Phi^\top$, and the tissue-specific covariance matrix $\Sigma_{\Lambda_s} = \Lambda_s \Lambda_s^\top$. In this particular settings, we do not have any identification problems.

In the genomic framework, it is crucial to assume sparsity. In fact, sparsity removes some entries in the loading matrix, resulting in network of genes highly expressed in one factor while the others equals to zero. Also, in settings where there are more variables than sample, the $p \gg n_s$, $s = 1, \dots, S$, it is fundamental to assume regularization on the loading matrix, such as prior or penalty. Therefore, we use the shrinkage prior of [1]. A relevant characteristic of this prior is that the shrinkage increases with the column index of the factor loading matrices. Considering additional factors that do not explain more than the one estimated result in a null weight in the matrix.

In the BMSFA setting, the prior for each elements of the common factor loading matrix Φ is

$$\begin{aligned} \phi_{pk} \mid \omega_{pk}, \tau_k &\sim N(0, \omega_{pk}^{-1} \tau_k^{-1}), \quad p = 1, \dots, P, k = 1, \dots, \infty, \\ \omega_{pk} &\sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \quad \tau_k = \prod_{l=1}^k \delta_l \quad \delta_1 \sim \Gamma(a_1, 1) \quad \delta_l \sim \Gamma(a_2, 1), \quad l \geq 2 \end{aligned}$$

with independent δ_l ($l = 1, 2, \dots$), a global shrinkage parameter τ_k for the k -th column of the factor loading matrix, and a local shrinkage parameter ω_{pk} for the element p in column k . The same prior is assumed for each element of the study-specific factor loading matrix Λ_s :

$$\lambda_{pj_s} \mid \omega_{pj_s}^s, \tau_{j_s}^s \sim N(0, \omega_{pj_s}^{s-1} \tau_{j_s}^{s-1}), \quad p = 1, \dots, P, \quad j_s = 1, \dots, \infty, \quad \text{and } s = 1, \dots, S,$$

$$\omega_{pj_s}^s \sim \Gamma\left(\frac{\mathbf{v}^s}{2}, \frac{\mathbf{v}^s}{2}\right) \quad \tau_{j_s}^s = \prod_{l=1}^{j_s} \delta_l^s \quad \delta_1^s \sim \Gamma(a_1^s, 1) \quad \delta_l^s \sim \Gamma(a_2^s, 1), \quad l \geq 2$$

with independent δ_l^s ($l = 1, 2, \dots$), a global shrinkage parameter $\tau_{j_s}^s$ for the j_s column of the factor loading matrix, and a local shrinkage parameter $\omega_{pj_s}^s$ for the element p in column j_s .

Finally, we consider an inverse gamma prior $\psi_{ps}^{-1} \sim \Gamma(a_\psi, b_\psi)$, for each element of the error covariance matrices ψ_{ps} , $p = 1, \dots, P$ $s = 1, \dots, S$. Posterior computation is achieved leveraging all the steps of the Gibbs sampler described in [3], specifically running 30000 chains with a burn-in of 10000. This MCMC framework provides good convergence and mixing.

3 Application in the GTEx Consortium

The GTEx consortium provides gene expression measures across 44 different tissues. In this particular application, we focus on 13 different and highly correlated brain tissues subtypes measured in 449 individuals from the GTEx v6p release. Brain regions are crucial to understand specific mechanism and to highlight gene-network and their associations with disease, such as Alzheimer disease. In fact, brain region study reveals that changes in behavior result in changes in gene expression in the brain [12]. We adopt the BMSFA in the GTEx consortium. We are interested in estimating brain tissue specific and remove the communalities that lack tissue heterogeneity. Particularly we are interested in detecting tissue-specific networking deriving from the tissue-specific covariance matrix $\Sigma_{\Lambda_s} = \Lambda_s \Lambda_s^\top$.

We represent the tissue network, by choosing two critical brain regions, i.e. the Basal Ganglia (Figure 1a) and the Hippocampus (Figure 1b). Specifically, the Basal Ganglia is associated with emotion and movement, while the hippocampus is associated to both long and short memory and in fact it is the first region to be impacted by Alzheimer disease. Note that the BMSFA is applied to all the 12 brain regions considered in the GTEx consortium, here for simplicity we report the network of these two relevant brain regions. The co-expression network is generated by the brain tissue specific part of covariance matrix, i.e. $\Sigma_{\Lambda_s} = \Lambda_s \Lambda_s^\top$. The truncation level for the graph is setting to 0.5, specifically we add edges between two genes if their brain tissue-specific correlation is greater than 0.5 (in absolute value).

We then proceed by performing the gene-set enrichment analysis to both the brain regions considered. Gene-set enrichment analysis is a powerful tool to detect if each

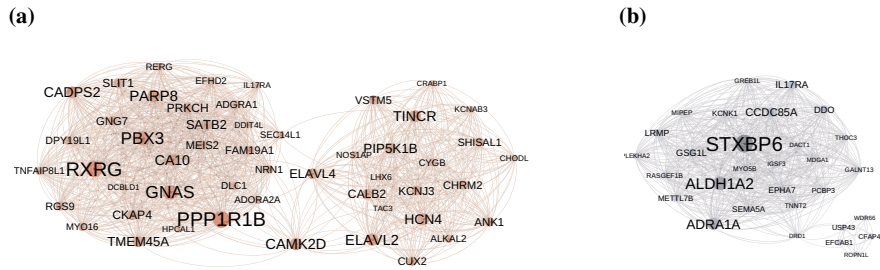


Fig. 1: Tissue-specific gene co-expression network of two brain regions, Basal Ganglia (a) and Hippocampus (b), considered in the analysis. We add edges between two genes if their brain tissue-specific correlation is greater than 0.5, in absolute value.

of the factor loading matrix is significantly associated/ enriched in a known biological pathway. We adopt the package `RTopper` in R in `Bioconductor` including all the gene sets corresponding to the pathways from `reactome.org`.

As a result, both the Basal Ganglia and the Hippocampus specific networks are significantly associated with the Axon Guidance Pathway. This pathway is crucially related to Pancreatic cancer and Parkinsons disease [2]. Just the Basal Ganglia specific network is significantly associated with Signaling by GPCR, an important pathway that regulates many different body's activities, such as hormones in the pancreas [8] and sleeps rhythms [11]. Finally, just the Hippocampus specific network is significantly associated with the Muscle Contraction and Calcium Pathway: two crucial pathways related to the Parkinson's and Alzheimer's diseases.

In conclusion, this analysis illustrates important characteristics of the approach, including its ability to capture biological signal coming from different brain region, and at the same time to isolate the source of variation arising from the common signal.

References

1. Bhattacharya, A., Dunson, D. B.: Sparse Bayesian infinite factor models, *Biometrika*, 291-306 (2011)
2. Ceyhan, G. O., et al.: The neurotrophic factor artemin promotes pancreatic cancer invasion. *Annals of surgery* **244**, 274-281 (2006)
3. De Vito, R., Bellio, R., Trippa, L., Parmigiani, G.: Bayesian Multi-study Factor Analysis for High-throughput Biological Data. *arXiv preprint arXiv:1806.09896* (2018)
4. De Vito, R., Bellio, R., Trippa, L., Parmigiani, G.: Multi-study Factor Analysis. *Biometrics* (2018), <https://doi.org/10.1111/biom.12974>
5. Dezső, Z., et al.: A comprehensive functional analysis of tissue specificity of human gene expression. *BMC biology* **6**, 49-57 (2008)
6. Greene, C., et al.: Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* **47**, 569-574 (2015)
7. GTEx Consortium: The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660 (2015)

8. Husted, A. S., et al.: GPCR-mediated signaling of metabolites. *Cell Metabolism* **25**, 777-796 (2017)
9. Kadota, K., et al.: Detection of genes with tissue-specific expression patterns using Akaike's Information Criterion (AIC) procedure. *Physiological genomics* **12**, 251-259 (2003)
10. Kouadjo, K. E., et al.: Housekeeping and tissue-specific genes in mouse tissues. *BMC genomics* **8**, 127-138 (2007)
11. Tsuneki, H., Sasaoka, T., Sakurai, T.: Sleep control, GPCRs, and glucose metabolism. *Trends in Endocrinology & Metabolism* **27**, 633-642 (2016)
12. Whitfield, C. W., Cziko, A., Robinson, G. E.: Gene expression profiles in the brain predict behavior in individual honey bees, *Science* **302**, 296-299 (2003)

Bayesian estimate of population count with false captures: a latent class approach

Stima Bayesiana della popolazione con false catture: un approccio basato sulle classi latenti

Davide Di Cecco, Marco Di Zio and Brunero Liseo

Abstract We propose a capture–recapture model for estimating the size of a population based on multiple lists in presence of out-of-scope units (false captures). Our Bayesian approach makes use of a class of log–linear models with a latent structure. We also address the presence of sources providing partial information implementing a Gibbs Sampler algorithm which generates a sample from the posterior distribution of the population size in the presence of missing data. The proposed method is applied to simulated data sets.

Abstract Si illustra un metodo di stima per la numerosità di una popolazione basato su liste multiple in presenza di unità erroneamente conteggiate (false capture). L'approccio bayesiano impiega una sotto classe di modelli log–lineari con struttura latente. Per utilizzare correttamente liste con informazioni parziali, si implementa un algoritmo di tipo Gibbs che genera un campione dalla distribuzione a posteriori del conteggio della popolazione in presenza di dati mancanti. La validità del metodo è verificata attraverso alcune simulazioni.

Key words: Bayesian Analysis, Capture–Recapture, Latent Class

1 Introduction

Estimating the size of a population is a central issue in several different fields. Many problems arise in practical applications, because the available information - either

Davide Di Cecco
ISTAT, Rome, e-mail: dicecco@istat.it

Marco Di Zio
ISTAT, Rome, e-mail: dizio@istat.it

Brunero Liseo
Sapienza Università di Roma e-mail: brunero.liseo@uniroma1.it

“capture occasions” in animal abundance problems or “lists” in multiple record systems context - often suffer from *i*) undercoverage, that is, some members of a population have zero chance of being captured or included in some of the lists; *ii*) overcoverage, that is, some units which do not belong to our population are erroneously included (“false captures”).

In this note we propose a capture–recapture model which takes into account these two aspects.

There are various proposals in the literature which adopt a complex model to deal with false capture in multiple lists. This methodology is frequent in animal abundance problems: see for example [2], [6], [12]. However, in these works false captures and misidentification are essentially ascribed to record linkage errors. In this work, we want to address a larger class of problems, where false captures do not have an identified source of error. Usually, possible causes of errors are inherently specific to the data at hand and require ad hoc procedures. Obviously, all the available information should be included in the process of identifying the erroneous captures. Ideally, recognizing and deleting spurious cases should constitute a first step of the analysis, followed by a capture–recapture technique on the “cleaned” data. However, the available information is often not sufficient to single out every false capture, and there might remain some uncertainty, difficult to deal with.

To our knowledge, the only contributions dealing with false captures with no restrictive hypothesis on the source of errors in multiple record systems are [8] and [5]. While the former proposes a standard Bayesian log-linear model, the latter extends the approach to include latent variables. However, in both cases, just one list is assumed to suffer from overcoverage.

The problem of overcoverage has received some interest in official statistics, where it is particular relevant when considering administrative sources to estimate population counts. In this context in fact, many possible sources of errors are recognized: record linkage errors are commonly addressed (see e.g., [11]); delays in record updating can lead to erroneous classification of units; more generally the differences in scope of the various sources are hard to harmonize correctly. At present, the most popular approach in official statistics provides for a Dual System Estimator (DSE). All pertinent sources are integrated into a unique population statistical register which is coupled with a survey. An estimate of the overcoverage rate is obtained through the comparison between the (supposedly) error free survey and the register via some supervised model. Then, the estimate is used to “correct” the DSE in some way. The advantage of the procedure is that the DSE is remarkably robust (see, e.g., [1]), and there is no need to rely on any complex model specification.

In our approach, we look at a multiple lists context, in order to exploit the information redundancy. In this context, a series of methodological issues may arise:

- non independence of the captures: it is important to consider potential dependencies among the various lists;
- whereas DSE is robust with respect to violation of basic hypotheses (e.g., the homogeneity of capture probabilities), this is not generally true in Multiple Record Systems;

- some units may have zero probability of being captured in some lists (e.g., lists targeting only specific subpopulation or different periods of time).

The proposal we discuss here relies on the following model assumptions: all possible erroneous captures are defined as random classification errors under a binary model. We assume the presence of two subpopulations: one comprising the out-of-scope units, and the other the in-scope units. Then, a two-component latent class model is assumed to describe the data. In order to model possible dependencies among captures of the same individual in different sources, we relax the classic conditional independence assumption of latent class models and we assume a general log-linear model. In order to deal with the structural absence of some subpopulations from specific lists, we propose to treat the uncappable units as missing information and develop a missing data approach. This idea has been already proposed in [4]. Here we present its Bayesian counterpart.

2 The model

Assume k lists or capture occasions are available, and let Y_i be the random variable indicating whether a unit is included in the i -th list, $i = 1, \dots, k$ (i.e., has been captured in the i -th occasion):

$$Y_i = \begin{cases} 1 & \text{if a unit is captured in the } i\text{-th list;} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbf{Y} = (Y_1, \dots, Y_k)$ denote the capture profile of a unit, and let $\{P(\mathbf{Y} = \mathbf{y}) = p_{\mathbf{y}}\}_{\mathbf{y} \in \{0,1\}^k}$ be the associated probability distribution.

Let $U(i)$ be the set of units that are catchable by list i , and let U be $\bigcup_i U(i)$. Let U_1 be our target population, with $U_1 \subset U$. The cardinality of U is N , the one of U_1 is N_1 . Let X be the latent variable identifying the units belonging to our target population:

$$X = \begin{cases} 1 & \text{if a unit belongs to } U_1; \\ 0 & \text{otherwise.} \end{cases}$$

Let $n_{\mathbf{y}}$ be the number of units having capture profile \mathbf{y} , of which $n_{x,\mathbf{y}}$ belong to the latent class x so that $n_{0,\mathbf{y}} + n_{1,\mathbf{y}} = n_{\mathbf{y}}$. The total number of observed unit is n_{obs} , while the units having capture history $\mathbf{y} = \mathbf{0} = (0, \dots, 0)$ are unobserved, so that $\sum_{\mathbf{y} \neq \mathbf{0}} n_{\mathbf{y}} = n_{obs}$, and $N = n_{obs} + n_{\mathbf{0}}$. Note that $n_{1,\mathbf{0}}$ is the number of units in U_1 that are not captured, while $n_{0,\mathbf{0}}$ is the number of uncaptured units which are in U but not in U_1 . We are interested in estimating $N_1 = \sum_{\mathbf{y}} n_{1,\mathbf{y}}$.

The latent class model under the conditional independence assumption (CIA) can be equivalently expressed as the mixture model

$$p_{\mathbf{y}} = \sum_{x=0,1} p_x \prod_{i=1}^k p_{y_i|x}, \quad (1)$$

where $p_{y_i|x}$ indicates the conditional probability $P(Y_i = y_i|X = x)$, or as in the log-linear model notation

$$[XY_1][XY_2] \cdots [XY_k], \quad (2)$$

which reports only the higher order interactions (generators) of the model. Any additional interaction term in (2) represents a relaxation of the CIA.

2.1 Prior distributions

The usual priors for log-linear models are based on Multivariate Gaussian distributions. Here we propose a different prior based on Dirichlet distributions. We find this approach easier in terms of elicitation of prior knowledge, and also from a computational point of view, since it allows us to develop a Gibbs sampler for obtaining a sample from the posterior distribution of N_1 , so avoiding the use of a Metropolis–Hastings algorithm.

To illustrate our proposal we start with decomposable models. In this case the prior distribution is simply the product of Dirichlet densities. In [3] it has been demonstrated that, if G is the dependence graph of the decomposable model, $\{\mathcal{C}_1, \dots, \mathcal{C}_g\}$ are the maximal cliques of G , and $(\mathcal{S}_2, \dots, \mathcal{S}_g)$ are defined as

$$\mathcal{S}_i = \mathcal{C}_i \cap \bigcup_{j=1}^{i-1} \mathcal{C}_j \quad i = 2, \dots, g,$$

the joint distribution can be written as the product of conditional distributions:

$$p_G = \prod_{i=1}^g p_{\mathcal{C}_i} \left(\prod_{j=2}^g p_{\mathcal{S}_j} \right)^{-1} = p_{\mathcal{C}_1} \prod_{i=2}^g p_{\mathcal{C}_i | \mathcal{S}_i}, \quad (3)$$

where p over a (sub)graph is the (marginal) distribution over the variables included in the (sub)graph. Let Θ be the vector of parameters $\Theta = (P_{\mathcal{C}_1}, P_{\mathcal{C}_2 | \mathcal{S}_2}, \dots, P_{\mathcal{C}_g | \mathcal{S}_g})$. We define a prior distribution on Θ as follows: for each $P_{\mathcal{C}_i | \mathcal{S}_i}$ and for each value of \mathcal{S}_i we set a Dirichlet distribution defined for each possible combination of values $\mathbf{y}_{\mathcal{C}} \in \{0, 1\}^{|\mathcal{C}|}$ of the variables in \mathcal{C} . The Dirichlet densities are independent by construction, and this class of priors is conjugate to (3).

In the case of a general log-linear model, we made use of the “Bayesian iterative proportional fitting” described in [10] in order to sample from a “Constrained Dirichlet”. That is, we generate samples from a Dirichlet distribution which satisfies the constraints given by the log-linear model. This prior has been rarely utilized in literature, and, as far as we know, has never been utilized in capture–recapture analysis.

Regarding N , in accordance with the literature on Bayesian capture–recapture, sensible options include:

- i) Jeffreys’ prior, i.e. $\pi(N) \propto 1/N$;

- ii) a hierarchical Poisson prior: $N \sim Poi(\lambda)$, $\lambda \sim Gamma(\alpha, \beta)$;
 - iii) Rissanen's prior ([9]), $\pi(N) \propto 2^{-\log^*(N)}$, where $\log^*(N)$ is the sum of the positive terms in the sequence $\{\log_2(N), \log_2(\log_2(N)), \dots\}$.
- We further assume that N and Θ are a priori independent.

3 Missing data

We propose a strategy useful to properly include sources which do not operate over certain subpopulations ("incomplete lists"). In fact, if we treat the uncachable units as sampling zeros, the final population size estimate would be biased.

The idea is to treat the incomplete lists as Missing at Random (MAR) information, i.e. assuming that, if they could operate on the whole population, they would retain the same joint distribution as in the observed subpopulations. In addition, we assume that we can distinguish whether a unit has not been captured in a list by chance or because it is out of the scope of that list, i.e., we can divide the population in strata where different set of lists operates. Then, certain profiles of the captured units are considered as partially observed, and we develop a data augmentation algorithm that imputes the complete capture histories using the rest of the data given the model.

We distinguish completely observed capture profiles, \mathbf{y} , from the partially observed capture profiles \mathbf{y}_{mis} . In addition, for each stratum, we have a structural zero \mathbf{z} consisting in a different combination of zeros and missing values. For example, in a 4-lists scenario with 2 strata, one where all lists operate and one where the first list does not operate, we have the structural zero $n_{0,0,0,0}$ in the first strata, and $n_{*,0,0,0}$ in the second, where the asterisk denotes the missing information.

Then, our Gibbs algorithm at iteration $t + 1$ has the following steps:

- 1) we sample the components of $\Theta^{(t+1)}$ from their posterior conditional Dirichlet distributions (constrained or not);
- 2) for each observed \mathbf{y} and \mathbf{y}_{mis} , we randomly divide all the observed values $n_{\mathbf{y}}$ and $n_{\mathbf{y}_{mis}}$ into the corresponding consistent complete sequences $n_{x,\mathbf{y}}$ according to their conditional probabilities;
- 3) if we adopt $\pi(N) \propto 1/N$, it has been demonstrated in [7] that we can sample all structural zero cells counts $n_{\mathbf{z}}$ from a Negative Multinomial distribution. Otherwise, if we choose an informative prior for N , we can use a Metropolis-Hasting step to generate a value for $N^{(t+1)}$ and then conditionally sample the structural zero cells such that $\sum_{\mathbf{z}} n_{\mathbf{z}} = N - n_{obs}$;
- 4) for each generated $n_{\mathbf{z}}$, we sample all complete sequences $n_{x,\mathbf{y}}$ consistent with \mathbf{z} .

4 Simulations

We compare the performance of our procedure in absence and in presence of missing data, and in absence and in presence of prior information. We report the results of a simulation for empirically assessing the proposed algorithm in the various settings. We considered 5 lists, A, B, C, D, E , and defined two scenarios: in the former, all five sources operate on the whole population, in the latter, there are three strata: one with all sources, another one where 4 out of five sources operate, and one where just three sources operate.

To evaluate the sensitivity to the prior distributions, we considered two additional cases: in the first, we used non informative priors for all parameters (all Dirichlet parameters equal to 1 and $\pi(N) \propto 1/N$), in the second we mimicked an informative context coming from an audit sample: we took a 5% sample of the generated complete population $[XABCDE]$, and fix the parameters of the Dirichlet prior equal to the observed counts in that sample.

We set $N = 10000$ and a proportion of out-of-scope units (both captured and non-captured) equal to 40%, so that the desired total N_1 is 6000 in expectation. We generated 500 independent samples from 2 models:

Model 1: $[XA][XB][XC][XD][XE]$, and Model 2: $[XABC][XD][XE]$.

For each sample, we registered the generated (“true”) values of N_1 (the target population size), and derived the marginal “observed” counts by omitting the structural zero cells. The model parameters have been set in such a way that the proportion of unobserved units (both in-scope and out-of-scope) is 20% in the scenario without missing information and about 30% in the scenario with three strata. For each simulation we calculated the posterior mean and the 95% credibility interval for N_1 .

The results are summarized in terms of sample bias in Table 1, and according to the average width of the 95% credibility intervals in Table 2.

Table 1 Results of the simulations. Sample Bias over the 500 samples in the four scenarios.

	Flat prior		Informative prior	
	No missing	Missings	No missing	Missings
Model 1	4.1	-8.3	3.6	5.1
Model 2	3.1	-12.5	-4.6	6.2

Model selection is a critical issue for capture–recapture modeling as population size estimate can be sensitive to changes in the parameterization. To have a hint on the robustness of the procedure under misspecification of the model, in the scenario described above with three strata, we generated a sample from model $[XABC][XD][XE]$, and estimated N_1 under two different models: model

Table 2 Results of the simulations. Average width of the 95% credibility intervals over the 500 samples in the four scenarios.

	Flat prior		Informative prior	
	No missing	Missings	No missing	Missings
Model 1	309.6	531.3	258.6	378.9
Model 2	185.2	345.6	160.2	263.8

$[XA][XB][XC][XD][XE]$, and the (non – decomposable) model including all 15 second order interactions but no higher order parameters. Results regarding the second model can be viewed in Figure 1, where one sees that the true value of N_1 is comprised in the 95% credibility interval, despite 5 parameters are missing (those relative to $[ABC]$, $[XAB]$, $[XAC]$, $[XBC]$ and $[XABC]$).

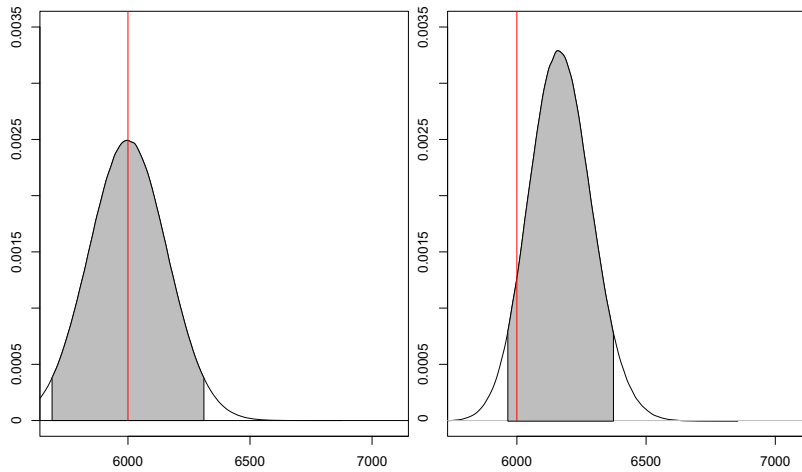


Fig. 1 Posterior distributions of N_1 under the generating model $[XABC][XD][XE]$ (left) and under the all–second–order–interactions model (right). The orange line indicates the true value of N_1 , the gray area the 95% HPD.

On the converse, the left panel of Figure 2 shows that the estimated posterior distribution of N_1 under the CIA model is far from the real value. To evaluate the influence of the prior distributions to compensate for the model misspecification, we set an informative prior as in the previous simulation, where a 5% audit sample establishes the Dirichlet priors parameters. As one can see in the right panel of Figure 2, even though informative priors influence the posterior in the right direction, their contribution seems insufficient to even include the true value of N_1 in the credibility interval.

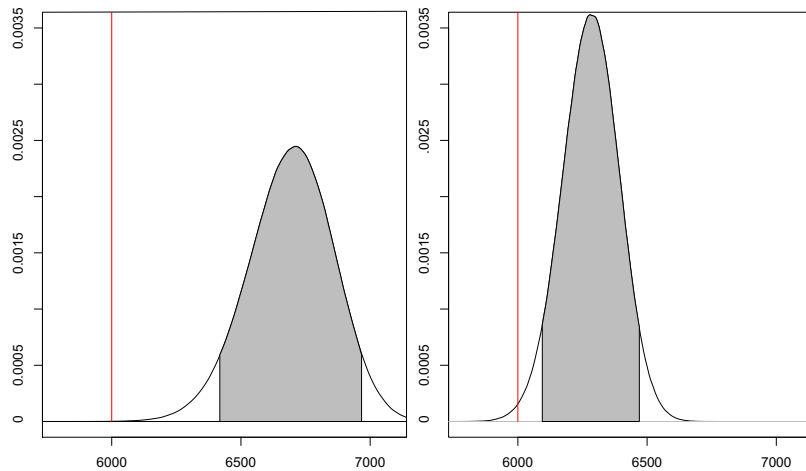


Fig. 2 Posterior distributions of N_1 under the CIA model, flat priors(left) and informative priors (right). The orange line indicates the true value of N_1 , the gray area the 95% HPD.

References

1. A. Chao, P.K. Tsay, S.H. Lin, W. Shau, and D. Chao. The applications of capture-recapture models to epidemiological data. *Statistics in medicine*, 20(20):3123–3157, 2001.
2. C. Q. da Silva. Bayesian analysis to correct false-negative errors in capture–recapture photo-ID abundance estimates. *Brazilian Journal of Probability and Statistics*, 23(1):36–48, 2009.
3. A. P. Dawid and S. L. Lauritzen. Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317, 1993.
4. D. Di Cecco, M. Di Zio, D. Filipponi, and I. Rocchetti. Population size estimation using multiple incomplete lists with overcoverage. *J. Off. Stat.*, 34(2):557–572, 2018.
5. D. A. Fegatelli, A. Farcomeni, and L. Tardella. Bayesian population size estimation with censored counts. In *Capture-Recapture Methods for the Social and Medical Sciences*, pages 371–385. Chapman and Hall/CRC, 2017.
6. W. A. Link, J. Yoshizaki, L. L. Bailey, and K. H. Pollock. Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. *Biometrics*, 66(1):178–185, 2010.
7. D. Manrique-Vallier and J. P. Reiter. Bayesian estimation of discrete multivariate latent structure models with structural zeros. *Journal of Computational and Graphical Statistics*, 23(4):1061–1079, 2014.
8. A. M. Overstall, R. King, S. M. Bird, S. J. Hutchinson, and G. Hay. Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. *Statistics in medicine*, 33(9):1564–1579, 2014.
9. J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of statistics*, 11(2):416–431, 1983.
10. J. L. Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.
11. A. Tancredi and B. Liseo. A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5(2B):1553–1585, 2011.
12. J. A. Wright, R. J. Barker, M. R. Schofield, A. C. Frantz, A. E. Byrom, and D. M. Gleeson. Incorporating genotype uncertainty into mark–recapture-type models for estimating abundance using DNA samples. *Biometrics*, 65(3):833–840, 2009.

Spherical regression with local rotations and implementation in R

Regressione sferica con rotazioni locali ed implementazione in R

Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor

Abstract We discuss the problem of estimating a nonparametric regression function when both the predictor and the response lie on a spherical domain, using local rotations. We also present the R package **nprotreg** [10], devoted to its implementation.

Abstract *Discutiamo il problema di stimare nonparametricamente una funzione di regressione nel caso in cui sia il predittore che la variabile risposta giacciono su un dominio sferico. Inoltre, presentiamo il pacchetto R **nprotreg** [10] che permette di implementare il metodo proposto.*

Key words: Local Rotations, Nonparametric Regression, **nprotreg** Package, Spherical Kernel

1 Introduction

Directional data arise in many scientific fields where observations are recorded as directions or angles relative to a fixed reference point. The unit hypersphere is the space of all directions, and is denoted by $\mathbb{S}^{d-1} := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$, $d \geq 2$, i.e. a $(d - 1)$ -dimensional manifold embedded in \mathbb{R}^d . The special cases $d = 2$ and $d = 3$ respectively refer to the unit circle and the *ordinary* sphere. Examples of such data include directions of winds, animals migrations, marine currents, Earth's main

Marco Di Marzio
DMQTE, Università di Chieti-Pescara. e-mail: marco.dimarzio@unich.it

Stefania Fensore
DMQTE, Università di Chieti-Pescara. e-mail: stefania.fensore@unich.it

Agnese Panzera
DiSIA, Università di Firenze. e-mail: a.panzera@disia.unifi.it

Charles C. Taylor
Department of Statistics, University of Leeds. e-mail: charles@maths.leeds.ac.uk

magnetic field, rocks fractures, etc. Because of the nonlinear nature of the manifold, all the statistical methods for dealing with directional data require adaptation. For comprehensive accounts of statistics for directional data see, for example, [8] and [7].

In this paper we discuss nonparametric regression for spherical data proposed by [4] along with the R package **nprotreg** [9] devoted to the implementation of this method. This is the case of multivariate regression where both the predictor and the response lie on \mathbb{S}^{d-1} . Dependence involving spherical variables, recently, has revealed to be a very interesting problem, and many examples apply in geology, crystallography, computer vision, shape analysis, etc.

Spherical regression has been introduced in the parametric framework by [2], who studied the estimation process of the unknown rotation relating the two variates lying on \mathbb{S}^{d-1} . Then, a nonparametric approach has been proposed by [3], where the regression function is locally modelled by using a polynomial. To predict the response they work component-wise, with the two following caveats: *i*) they exclude any correlation between dimensions; *ii*) they do not require both the variates to lie on spheres of the same dimension. Specifically, they proposed a local polynomial fitting multi-output approach, in the sense that each coordinate of the response variable is treated with a distinct regression.

The novelty of the model proposed by [4], which is discussed in this paper, is the flexibility of the regression function due to local rotations, i.e. for each location of the manifold a specific rotation matrix has to be estimated. In such a framework the shape of the scatter of predicted values is, in general, different from the shape of predictors; such a transformation is usually referred to as a *non-rigid* rotation. Moreover this scenario is set up as *simple regression* – although the observations are multidimensional – and the need to model correlation is removed.

The paper is organized as follows. In Section 2 we recall some basic facts about the parametrization of points on the sphere. Section 3 deals with the regression model and the resulting estimators. Finally, in Section 4 we present the **nprotreg** package along with an illustrative example.

2 Mathematical tools

According to the *tangent-normal* decomposition, given a fixed $\mathbf{x} \in \mathbb{S}^{d-1}$, any vector $\mathbf{u} \in \mathbb{S}^{d-1}$ can be expressed as $\mathbf{u}(\boldsymbol{\xi}, \theta) := \mathbf{x} \cos(\theta) + \boldsymbol{\xi} \sin(\theta)$, where θ is the angle between \mathbf{u} and \mathbf{x} , and $\boldsymbol{\xi}$ is a vector orthogonal to \mathbf{x} . Now, for a function $g : \mathbb{S}^{d-1} \rightarrow \mathbb{R}$, let $\bar{g}(\mathbf{x}) := g(\mathbf{x}/\|\mathbf{x}\|)$ be the homogeneous extension of g to $\mathbb{R}^d \setminus \{\mathbf{0}_d\}$, with $\mathbf{0}_d$ being the d -dimensional zero vector. The directional derivative of order ℓ of \bar{g} at \mathbf{x} in the direction of $\boldsymbol{\xi}$, denoted by $\mathcal{D}_{\boldsymbol{\xi}}^{(\ell)} \bar{g}(\mathbf{x})$, is $\left. \frac{\partial^\ell}{\partial \theta^\ell} g(\mathbf{u}(\boldsymbol{\xi}, \theta)) \right|_{\theta=0} = \boldsymbol{\xi}' \nabla_{\bar{g}(\mathbf{x})}^\ell \boldsymbol{\xi}^{\otimes(\ell-1)}$, where $\nabla_{\bar{g}(\mathbf{x})}^\ell$ denotes the matrix of the derivatives of order ℓ of \bar{g} at \mathbf{x} , and, for a vector \mathbf{a} , $\mathbf{a}^{\otimes \ell}$ stands for its Kroneckerian power of order ℓ . Then, under suitable continuity assumptions, a p th-order Taylor series expansion of g around \mathbf{x} yields

$$g(\mathbf{u}) \approx g(\mathbf{x}) + \sum_{\ell=1}^p \frac{\theta^\ell}{\ell!} \boldsymbol{\xi}' \nabla_{\bar{\mathbf{g}}(\mathbf{x})}^\ell \boldsymbol{\xi}^{\otimes(\ell-1)}. \quad (1)$$

We also recall the formula of the matrix exponential. For any matrix \mathbf{A} of order d , we have $\exp(\mathbf{A}) = \mathbf{I}_d + \mathbf{A} + \mathbf{A}^2/2 + \dots$, with \mathbf{I}_d denoting the identity matrix of order d . Given $\mathbf{x} \in \mathbb{S}^{d-1}$, any rotation matrix $\mathbf{R}_{\mathbf{x}}$ has an exponential form $\mathbf{R}_{\mathbf{x}} = \exp(\mathbf{S}_{\mathbf{x}})$, where $\mathbf{S}_{\mathbf{x}}$ is a skew-symmetric matrix, i.e. $\mathbf{S}_{\mathbf{x}}^T = -\mathbf{S}_{\mathbf{x}}$. Real skew-symmetric matrices are mapped into the *Lie group* of orthogonal matrices by the matrix exponential. For further details regarding exponentials of skew-symmetric matrices see [6].

Provided that the homogeneous extension of each non-zero entry of the skew-symmetric matrix $\mathbf{S}_{\mathbf{x}_i}$, say $s_{jk}(\mathbf{x}_i)$, with $(j, k) \in (1, \dots, d) \times (1, \dots, d)$, with $j \neq k$, has p continuous derivatives in a neighbourhood of $\mathbf{x} \in \mathbb{S}^{d-1}$, expansion (1), for $s_{jk}(\mathbf{x}_i)$ around \mathbf{x} , yields the following local approximation for $\mathbf{R}_{\mathbf{x}_i}$

$$\mathbf{R}_{\mathbf{x}_i} = \exp(\mathbf{S}_{\mathbf{x}_i}) \approx \exp\left(\mathbf{S}_{\mathbf{x}} + \sum_{\ell=1}^p \mathbf{D}_{\mathbf{S}_{\mathbf{x}}}^{(\ell)}(\mathbf{x}_i, \mathbf{x})\right), \quad (2)$$

where $\mathbf{D}_{\mathbf{S}_{\mathbf{x}}}^{(\ell)}(\mathbf{x}_i, \mathbf{x})$ is the matrix of order d having $\theta_i^\ell / (\ell!) \mathcal{D}_{\xi_i}^{(\ell)} \bar{s}_{jk}(\mathbf{x})$ as its (j, k) th entry. A further approximation using the expansion of the matrix exponential yields

$$\mathbf{R}_{\mathbf{x}_i} \approx \mathbf{R}_{\mathbf{x}} \left(\mathbf{I}_d + \sum_{\ell=1}^p \mathbf{D}_{\mathbf{S}_{\mathbf{x}}}^{(\ell)}(\mathbf{x}_i, \mathbf{x}) \right). \quad (3)$$

3 The model

Consider a pair of random variables $(\mathcal{X}, \mathcal{Y})$, both taking values on \mathbb{S}^{d-1} , and assume that the regression of \mathcal{Y} on \mathcal{X} exists for each $\mathbf{x} \in \mathbb{S}^{d-1}$. Such a regression function can be written in a very flexible way by specifying a distinct rotation for each predictor value \mathbf{x}

$$\mathbb{E}[\mathcal{Y} \mid \mathcal{X} = \mathbf{x}] = \mathbf{R}_{\mathbf{x}}^T \mathbf{x}. \quad (4)$$

The experimental error is represented as a small random rotation of the true regression. Using the matrix exponential formula for the random error term $\boldsymbol{\varepsilon}$, the regression for independent copies $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)$ of $(\mathcal{X}, \mathcal{Y})$ can be written as

$$\mathbf{y}_i = \exp(\Phi(\boldsymbol{\varepsilon}_i)) \mathbf{R}_{\mathbf{x}_i}^T \mathbf{x}_i, \quad i \in (1, \dots, n), \quad (5)$$

where the function $\Phi(\mathbf{a})$ maps an \mathbb{R}^d vector $\mathbf{a} = (a_1, a_2, \dots, a_{d(d-1)/2})^T$ into a skew-symmetric matrix; for example, for $d = 3$ we could use

$$\Phi(\mathbf{a}) = \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix};$$

and the $\boldsymbol{\varepsilon}_i$ s satisfy $E[\boldsymbol{\varepsilon}_i | \mathbf{x}_i] = \mathbf{0}_d$, where $\mathbf{0}_d$ stands for a d -dimensional vector of zeros, and $\text{Var}[\mathbf{y}_i | \mathbf{x}_i] = \boldsymbol{\Sigma}_{\mathbf{x}_i}$, with $\boldsymbol{\Sigma}_{\mathbf{x}_i}$ being a matrix of order d with finite entries. Observe that, if $\boldsymbol{\varepsilon}_i$ has entries close to zero, then $\exp(\Phi(\boldsymbol{\varepsilon}_i)) \approx \mathbf{I}_d$. This is equivalent to assuming that the distribution of $\mathcal{Y} | \mathcal{X} = \mathbf{x}_i$ has expectation $\mathbf{R}_{\mathbf{x}_i}^T \mathbf{x}_i$.

Model (5) requires the estimation of a distinct rotation matrix corresponding to each \mathbf{x}_i and this makes the task computationally intensive. However, assuming that, within wide enough regions, the rotation matrices are “similar” makes the treatment less cumbersome. Assuming that (4) is a *smooth* mapping, i.e. $\lim_{\mathbf{x}_i \rightarrow \mathbf{x}} \mathbf{R}_{\mathbf{x}_i} = \mathbf{R}_{\mathbf{x}}$, in the sense that, if $\mathbf{x}_1 \approx \mathbf{x}_2$ then $\mathbf{R}_{\mathbf{x}_1} \approx \mathbf{R}_{\mathbf{x}_2}$ and so $\mathbf{y}_1 \approx \mathbf{y}_2$, guarantees a local approximation of $\mathbf{R}_{\mathbf{x}_i}$ in terms of $\mathbf{R}_{\mathbf{x}}$ under suitable smoothness conditions.

Therefore in Section 3.1 and Section 3.2 we approximate $\mathbf{R}_{\mathbf{x}_i}$ by (2), or, equivalently, by (3) respectively using $p = 0$ and $p = 1$. $\mathbf{R}_{\mathbf{x}}$ can be approximated by *all* the n rotations $\mathbf{R}_{\mathbf{x}_i}$, $i \in (1, \dots, n)$, more accurately as the location \mathbf{x}_i is closer to \mathbf{x} . Consequently, each observation \mathbf{x}_i can participate in the estimation of the rotation at \mathbf{x} with the caveat that its contribution needs to be smaller for larger $\|\mathbf{x}_i - \mathbf{x}\|$.

3.1 Local constant estimator

Using a one-term approximation in equation (3), to find $\hat{\mathbf{y}} | \mathbf{x} = \hat{\mathbf{R}}^T \mathbf{x}$ we can obtain $\hat{\mathbf{R}}$ as the solution of the following locally weighted least squares problem

$$\hat{\mathbf{R}}_{\mathbf{x}} = \underset{\mathbf{R}_{\mathbf{x}} \in \text{SO}(d)}{\text{argmin}} \sum_{i=1}^n \|\mathbf{y}_i - \mathbf{R}_{\mathbf{x}}^T \mathbf{x}_i\|^2 K_{\kappa}(\mathbf{x}_i^T \mathbf{x}), \quad (6)$$

where the weight function $K_{\kappa}(\mathbf{x}_i^T \mathbf{x})$ is a *spherical kernel* chosen to reflect the geodesic distance from \mathbf{x}_i to \mathbf{x} . A spherical kernel is a unimodal spherical density with rotational symmetry about its mean direction and concentration parameter $\kappa > 0$. The kernel function has the rôle of tuning the contribution of the observations in a neighbourhood of an estimation point \mathbf{x} and the κ controls the width of the neighbourhood of \mathbf{x} in so that smaller values of κ give wider neighbourhoods.

The solution to (6) is obtained by Singular Value Decomposition (SVD) of $\mathbf{Y}^T \mathbf{X}$, where \mathbf{X} and \mathbf{Y} are both $n \times d$ matrices and \mathbf{x}_i^T and \mathbf{y}_i^T are their respective i th rows. To preclude those solutions which include a reflection, we use $\hat{\mathbf{R}} = \mathbf{V} \boldsymbol{\Delta} \mathbf{U}^T$ in which \mathbf{U} and \mathbf{V} are obtained from the SVD: $\mathbf{Y}^T \mathbf{W}_{\kappa}(\mathbf{x}) \mathbf{X} = \mathbf{U} \boldsymbol{\Delta} \mathbf{V}^T$, with $\boldsymbol{\Delta}$ being a diagonal matrix of order d with entries $(1, \dots, 1, |\mathbf{V} \mathbf{U}^T|)$ and, for $\mathbf{a} \in \mathbb{S}^{d-1}$, $\mathbf{W}_{\kappa}(\mathbf{a})$ is a diagonal matrix of order n having $K_{\kappa}(\mathbf{x}_i^T \mathbf{a})$ as its (i, i) th entry.

3.2 Local linear estimator

Using a two-terms approximation in equation (3), the estimator $\hat{\mathbf{R}}$ is given by the solution of the following locally weighted least squares problem

$$\underset{\mathbf{R}_x \in \text{SO}(d), \mathbf{D}_{\mathbf{S}_x}^{(1)}(\mathbf{x}_i, \mathbf{x}) \in \text{Skew}_d}{\text{argmin}} \sum_{i=1}^n \left\| \mathbf{y}_i - \mathbf{R}_x^T \left\{ \mathbf{I}_d + \mathbf{D}_{\mathbf{S}_x}^{(1)}(\mathbf{x}_i, \mathbf{x}) \right\}^T \mathbf{x}_i \right\|^2 K_\kappa(\mathbf{x}_i^T \mathbf{x}) \quad (7)$$

where $\mathbf{D}_{\mathbf{S}_x}^{(1)}(\mathbf{x}_i, \mathbf{x})$ is a skew-symmetric matrix of directional derivatives. The number of terms to estimate are $d(d+1)(d-1)/2$ and there is no closed form solution. However, numerical procedures using nonlinear minimization seem quite stable for the case $d = 3$.

3.2.1 Iterations and algorithm

In order to minimize equation (7) for the case $d = 3$ we use the following algorithm. There are four rotation matrices and $(d+1)d^2 = 36$ parameters. Because any rotation is the matrix exponential of a skew-symmetric matrix, only 12 parameters need to be estimated. Now, given a design point \mathbf{x} and concentration parameter κ , let \mathbf{P} be a 3×4 matrix of the unknown parameters and set $\mathbf{M} = \mathbf{PD}$ with $\mathbf{D} = [\mathbf{1}_d \mathbf{x} \mathbf{1}_n^T - \mathbf{X}^T]$, and $\mathbf{1}_m$ being a vector of 1s of length m . For each $i = 1, \dots, n$ obtain the fitted \mathbf{y}_i :

$$\begin{aligned} \mathbf{S} &= \Phi(\mathbf{M}[i,]) \\ \theta &= \|\mathbf{M}[i,]\| \\ \mathbf{S} &= \mathbf{S}/\theta \\ \mathbf{R} &= \mathbf{I}_3 + \sin(\theta)\mathbf{S} + (1 - \cos(\theta))\mathbf{S}^2 \\ \hat{\mathbf{y}}_i &= \mathbf{R}^T \mathbf{X}[i,] \end{aligned}$$

after which we can compute the weighted sum as in equation (7). In the above loop the function Φ returns a skew symmetric matrix, and $\mathbf{X}[i,]$ and $\mathbf{X}[i,]$ denote vectors formed by the i th row and column of a matrix \mathbf{X} , respectively. We also use the Rodrigues' rotation formula due to [5] to express a rotation matrix in terms of an angle (θ) and an axis which can be determined from the elements of a skew symmetric matrix. The solution of equation (6) can be used to predict the response for a point \mathbf{x} , given \mathbf{X} , \mathbf{Y} and κ . This can be seen as a function f such that $\hat{\mathbf{y}}_{\mathbf{x}} = f(\mathbf{x}, \mathbf{X}, \mathbf{Y} \mid \kappa) = \hat{\mathbf{R}}_{\mathbf{x}}^T \mathbf{x}$, where $\hat{\mathbf{R}}_{\mathbf{x}}$ is the solution to (6). Given the nonrigid solution, the interpoint distances of estimated $\hat{\mathbf{y}}_{\mathbf{x}_i}$ will not be the same as the \mathbf{x}_i , which allows us to consider a Newton-Raphson-like iterative rotation fitting procedure.

4 nprotreg Package

In this section we introduce the **nprotreg** package that basically fits the sphere-sphere regression models discussed before by estimating locally weighted rotations. Also, it simulates sphere-sphere data according to non-rigid rotation models, including the rigid rotation as a particular case. A cross-validation procedure is provided

to select the smoothing parameters and a particular spherical kernel is proposed. The package is available from the Comprehensive R Archive Network at <http://CRAN.R-project.org/package=nprotreg>, see [9]. In the following we provide the complete list of functions available in the package, and, then, we describe the core ones.

All the functions available in the package are briefly described in Table 1. The content of the package is divided into two parts: the conversion functions which are self-explanatory, and the regression ones which make the user able to easily implement the methods discussed in Section 3.

Table 1 Functions available in **nprotreg** package.

Conversion functions	Description
<code>convert_cartesian_to_spherical</code>	Converts cartesian to spherical coordinates
<code>convert_spherical_to_cartesian</code>	Converts spherical to cartesian coordinates
Regression functions	Description
<code>get_skew_symmetric_matrix</code>	Gets a 3-by-3 skew symmetric matrix
<code>get_equally_spaced_points</code>	Generates equally spaced points on 3D sphere
<code>cross_validate_concentration</code>	Cross-validates the concentration parameter in a 3D spherical regression
<code>fit_regression</code>	Fits a 3D nonparametric spherical regression
<code>simulate_regression</code>	Simulates a 3D spherical regression
<code>simulate_rigid_regression</code>	Simulates a rigid 3D spherical regression
<code>weight_explanatory_points</code>	Weights the specified explanatory points in a 3D spherical regression

The main function of the package is `fit_regression()`. It returns spherical points obtained by locally rotating specified evaluation points, given an approximated model for local rotations and a weighting scheme for the observed data set. We now describe in detail the usage of this crucial function:

```
fit_regression(evaluation_points, explanatory_points, response_points,
  concentration, weights_generator = weight_explanatory_points,
  number_of_expansion_terms = 1, number_of_iterations = 1,
  allow_reflections = FALSE)
```

We require a first matrix of dimension $m \times d$ whose rows contain the points at which the regression will be estimated (`evaluation_points`), and other two matrices, \mathbf{X} (`explanatory_points`) and \mathbf{Y} (`response_points`), each of dimension $n \times d$, which respectively refer to the explanatory and response observations. The `concentration` value is a non negative scalar whose reciprocal is proportional to the Euclidean bandwidth, while `weights_generator` is a function that, given a matrix of m evaluation points, returns a $n \times m$ weight matrix (whose j th column contains the weights assigned to the explanatory points while analysing the j th evaluation point). The `number_of_expansion_terms` argument refers to the number of terms to be included in the expansion of the matrix exponential (3) to approximate a local rotation matrix; it must be 1 or 2. While, `number_of_iterations` refers to the number of rotation fitting steps to be executed discussed in Sec-

tion 3.2.1. At each step, the points estimated during the previous one are exploited as the current explanatory points. Finally, `allow_reflections` is a logical scalar value such that when set to `TRUE` reflections are allowed; it is ignored (`FALSE`) if `number_of_expansion_terms` is 2.

The matrices \mathbf{X} and \mathbf{Y} can either be read into R in the usual way, or can be generated by using the function `get_equally_spaced_points()`, which creates *approximately* equally spaced points on the unit sphere for $d = 3$.

The concentration parameter can be chosen by cross-validation using the function `cross_validate_concentration()` which minimizes $(nd)^{-1} \sum_{i=1}^n \|\hat{\mathbf{y}}_i^{(i)} - \mathbf{y}_i\|^2$, where $\hat{\mathbf{y}}_i^{(i)}$ is the prediction corresponding to \mathbf{x}_i , when the rotation matrix is estimated using all the data except the i th observation.

The function `weight_explanatory_points()` returns the weights assigned to the specified `explanatory_points` for each `evaluation_point`, given a concentration parameter κ . We use, as a weight function: $K_\kappa(\mathbf{x}_i^T \mathbf{x}) = \exp(\kappa(\mathbf{x}_i^T \mathbf{x} - 1))$, that is a rotationally symmetric function with maximum at \mathbf{x} , and a parameter κ which governs how much the weight spreads around \mathbf{x} . As a result, \mathbf{x}_i will receive a bigger weight the closer it is to \mathbf{x} , and for larger κ .

The function `simulate_regression()` returns the response points corresponding to the specified `explanatory_points`, given a model for local rotations (`local_rotation_composer`) – whose output is the vector of independent components of a skew symmetric matrix local to an explanatory point – and an error term sampler (`local_error_sampler`) – whose output is the vector of error term local to an explanatory point –.

Given its independent components, a skew symmetric matrix can be created using the function `get_skew_symmetric_matrix()`.

Finally, `simulate_rigid_regression()` returns the response points corresponding to the specified `explanatory_points`, but given a *rigid* rotation model (`rotation_matrix`) and an error term (`local_error_sampler`).

4.1 Illustrative example

After loading the library `nprotreg` we create the matrices of $n = 100$ explanatory (\mathbf{X}) and response points (\mathbf{Y}) using the functions `get_equally_spaced_points()` and `simulate_regression()`. Then, we select the concentration parameter by means of `cross_validate_concentration()` and, after setting the evaluation points equal to the explanatory ones, we fit the regression model getting $\hat{\mathbf{Y}}$.

```
R> library(nprotreg)
R> n <- 100
R> X <- get_equally_spaced_points(n)
R> evaluation_points <- X

R> local_rotation_composer <- function(point) {
+   independent_components <- (1 / 8) *
+   c(exp(2 * point[3]), - exp(2 * point[2]), exp(2 * point[1])) }

```

```

R> local_error_sampler <- function(point) {
+   noise <- rnorm(3, sd = 0.01) }

R> Y <- simulate_regression(explanatory_points = X,
+   local_rotation_composer, local_error_sampler)

R> k <- cross_validate_concentration(concentration_upper_bound = 100,
+   explanatory_points = X, response_points = Y)$concentration

R> Ys <- fit_regression(evaluation_points,
+   explanatory_points = X, response_points = Y, concentration = k)
R> Y_prediction <- Ys[[1]]$fitted_response_points

```

We can visualize the rotation models in a graphical 3-D plot using the package **rgl** [1] or projecting the sphere onto the plane, as shown in Fig. 1.

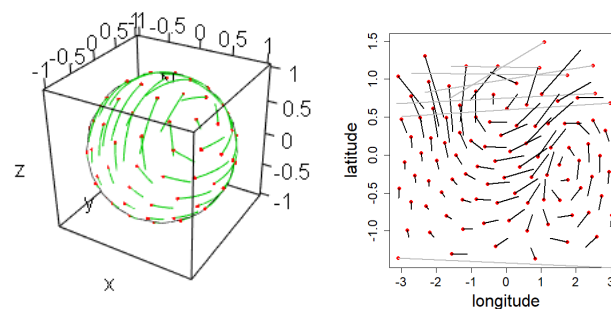


Fig. 1 Explanatory points \mathbf{X} (red) and fitted values $\hat{\mathbf{Y}}$ (left); Latitude-longitude projection (right).

References

1. Adler, D., Murdoch Dea: **rgl**: 3D Visualization Using OpenGL. URL <https://r-forge.r-project.org/projects/rgl/> (2018)
2. Chang, T.: Spherical regression. *The Annals of Statistics*, **14**, 907–924 (1986)
3. Di Marzio, M., Panzera, A., Taylor, C.C.: Nonparametric regression for spherical data. *J. Amer. Statist. Assoc.* **109**, 748–763 (2014)
4. Di Marzio, M., Panzera, A., Taylor, C.C.: Nonparametric rotations for sphere-sphere regression. *J. Amer. Statist. Assoc.* <http://dx.doi.org/10.1080/01621459.2017.1421542> (2018)
5. Euler, L.: *Nova Methodus Motum Corporum Rigidorum Determinandi*. *Novi Commentari Acad. Imp. Petrop.*, **20**, 208–238 (1775)
6. Gallier, J., Xu, D.: Computing exponentials of skew-symmetric matrices and logarithm of orthogonal matrices. *International Journal of Robotics and Automation*, **18**, 10–20 (2003)
7. Ley, C., Verdebout, T.: *Modern Directional Statistics*. Chapman & Hall/CRC Press, Boca Raton, Florida (2017)
8. Mardia, K.V., Jupp, P.E.: *Directional Statistics*. J. Wiley, Chichester (2008)
9. Taylor, C.C., Lafratta, G., Fensore, S.: **nprotreg**: Nonparametric Rotations for Sphere-Sphere Regression. URL <http://CRAN.R-project.org/package=nprotreg> (2018)
10. R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/> (2019)

A clustering method for network data to analyse association football playing styles

Un metodo di raggruppamento per dati di rete finalizzato all'analisi degli schemi di gioco nel calcio

Jacopo Diquigiovanni

Abstract We develop an innovative hierarchical clustering method to divide a sample of undirected weighted networks into clusters. The methodology consists of two phases: after an ad hoc preprocessing data phase that puts the networks in a sample context, in the second phase the method groups the networks on the basis of the similarity between the community structures of the processed networks. Starting from the representation of the association football team's playing style as a network, we apply the method to group the Italian Serie A teams' performances and consequently detect the main 15 tactics shown during the 2015–2016 season. The information obtained is used to verify the effect of the playing styles on the number of goals scored.

Abstract Il presente lavoro propone un nuovo metodo di raggruppamento gerarchico finalizzato a suddividere in gruppi un campione di reti pesate indirette. La metodologia consta di due fasi: dopo una fase di pre-elaborazione mirata ad inserire le singole reti in un contesto campionario, nella seconda fase queste vengono assegnate ai gruppi sulla base della similarità tra le strutture di comunità delle reti trasformate. Rappresentando lo schema di gioco di una squadra di calcio tramite una rete è possibile applicare il metodo per trovare i principali 15 schemi di gioco del campionato di Serie A stagione 2015-2016. L'informazione ottenuta è utilizzata per verificare se ed in che modo lo schema di gioco influisca sul numero di gol segnati.

Key words: community structures, Dixon and Coles model, hierarchical clustering, playing style analysis, population of networks

Jacopo Diquigiovanni

Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, IT-35121 Padova, Italy, e-mail: jacopo.diquigiovanni@phd.unipd.it

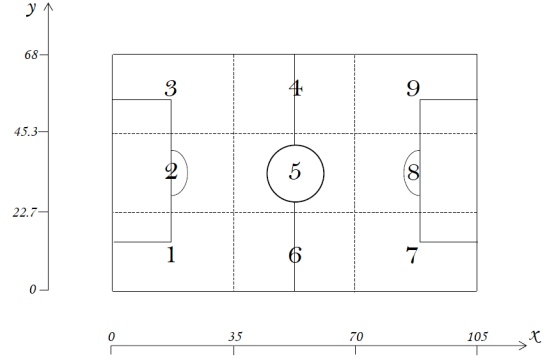
1 Introduction

The playing style (or tactic), defined as the way a team moves the ball on the pitch, is a feature of growing interest in association football (referred to simply as “football” hereafter). In statistical terms, the playing style could be properly represented by a network whose nodes are the team’s players and whose edges describe the movements of the ball between them (see, e.g., [6]); nevertheless, in the specific context of this work this representation is not convenient since it makes comparisons between different tactics not immediate as the players vary from squad to squad. In view of this, our approach considers the playing style as a weighted undirected network whose nodes are different areas of the football pitch - that is obviously shared between the teams - and whose edges describe the movements of the ball between these areas. Specifically, for the sake of simplicity the areas are obtained by dividing the length and the width of the football field into equal segments and the number of areas is set equal to 9 ($K = 9$) since it is associated with the best known subdivision of the length of the pitch (defense zone–midfield zone–attack zone) and with the best known subdivision of the width of the pitch (right zone–central zone–left zone). Naturally, this division of the pitch is heuristic and we plan to consider different values of K to verify how the results change and to evaluate the methodology when the areas have different shapes, e.g. considering the penalty area as a specific area. Considering a specific team playing a specific match, the data - provided by InStat and referring to the 380 matches of the Italian Serie A 2015-2016 season - detects spatial coordinates (x, y) of specific plays (passes, dribbling, tackles, shots) made by the team’s players during the game, with the x position along the length of the football field and the y position along the width of the football field. Starting from the spatial coordinates, each play is assigned to one of the nine areas, labeled by the number from 1 to 9 (see Figure 1). Therefore the weight of the edge linking node i and node j is represented by the number of pairs of consecutive plays made by the players of the team and which take place in areas i and j respectively. If an event is not preceded and followed by a teammate’s play, this is ignored as an isolated event; moreover, a self-loop is created when $i = j$. As each football game is made up of two distinct networks, the sample size is $n = 760$.

The work proposes a clustering method to group a sample of undirected weighted networks according to a specific criterion. The only constraint required is that the statistical units, *i.e.* the networks, share the set of nodes $\mathcal{K} = \{k_1, k_2, \dots, k_K\}$. The method, by clustering the performances of the teams as networks, detects the main 15 playing styles, and this information is used to investigate how a team’s tactic affects the number of goals scored by implementing an extension of the Dixon and Coles model [3]. We believe that this method may provide coaches and football journalists with a data-driven tool to evaluate playing styles’ efficacy in order to facilitate the opening up to ‘statistical culture’ by football operators.

The paper is organised as follows: in Sect. 2 the clustering method is developed, in Sect. 3 the method is applied to the analysis of playing styles and in Sect. 4 an overview of the main results is provided.

Fig. 1 Division of the pitch into nine areas. The areas are obtained by dividing the length and the width of the football field into three equal segments. The pitch measurements are 105×68 m and the teams attack from left to right.



2 The method

Focusing on a single network, the core of the cluster analysis is represented by the analysis of community structure, the aim of which is to find groups (communities) of nodes with many connections within communities and sparse connections between communities [5]. Considering the broader framework of this work, each network should be considered not only as a set of linked nodes, but also as a statistical unit drawn from a generic population of networks. In view of this, the method consists of two phases: after an ad hoc preprocessing data phase that puts the networks in a sample context (*first phase*), in the *second phase* the method groups the networks on the basis of the similarity between the community structures, detected by the widespread modularity-based approach known as Louvain method [1], of the processed networks.

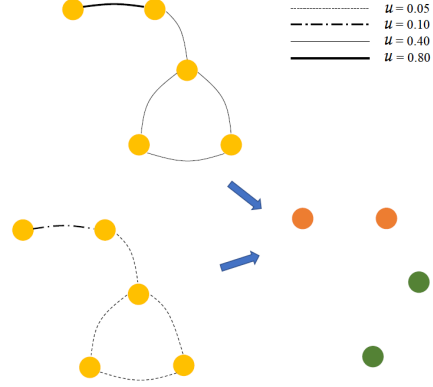
2.1 First phase

Let R be a generic network of a sample of n undirected weighted networks, with $\mathcal{K} = \{k_1, k_2, \dots, k_K\}$ the set of nodes shared between the networks. The procedure to obtain the processed network R' starting from R is described below. Let $w_{ij} = (w_{ij}^1, \dots, w_{ij}^n)$ be a vector with w_{ij}^t weight of the edge linking nodes i and j in t -th network ($t = 1, \dots, n$). The normalization of vector w_{ij} is defined as follows:

$$u_{ij}^t = \begin{cases} \frac{w_{ij}^t - a_{ij}}{b_{ij} - a_{ij}} \in [0, 1] & \text{if } a_{ij} \neq b_{ij}, \quad t = 1, \dots, n \\ 0.5 & \text{if } a_{ij} = b_{ij}, \quad t = 1, \dots, n \end{cases}$$

with $b_{ij} = \max(w_{ij}^1, \dots, w_{ij}^n)$ and $a_{ij} = \min(w_{ij}^1, \dots, w_{ij}^n)$. Thus a low value of u_{ij}^t highlights that the connection between the pair of nodes is numerically inferior com-

Fig. 2 Graphical representation of the problem in the application of Louvain method.



pared to that detected in the other networks, while a high value means the connection is numerically significant compared to that observed in the others. This first step is useful but not comprehensive because another issue is quite common in this context. Let us consider the following example: let $K = 5$ be the number of nodes and R_1 be a network in which all the normalized weights are $u_{1,b} = 0.05$ except for the edge linking the first two nodes which weighs $u_{1,a} = 0.10$ (see the bottom left of Figure 2). Let R_2 also be a network in which all the normalized weights are decidedly higher $u_{2,b} = 0.40$ and the edge linking the first two nodes weighs $u_{2,a} = 0.8$ (see the up of Figure 2). As shown by the bottom right of Figure 2, the Louvain method detects the same communities in the two normalized networks, a fact that implies that the two units have maximum similarity (if a measure of similarity that assigns maximum value when the partitions are identical is used). However in the first case the community including the first two nodes is found only because the other connections of the network are even smaller; evidently in this specific case detecting this community is counterproductive as $u_{1,a}$ is much lower than the other weights observed in the sample between the same pair of nodes. This critical issue is due to the fact that the Louvain method, by construction, considers the connections within the network regardless of the broader framework of this work.

In view of this, a feasible approach to limit the problem consists of introducing a threshold. Starting from the normalized weight u_{ij}^t we obtain the value $w_{i,j}^t$ defined as follows:

$$w_{i,j}^t = \begin{cases} u_{ij}^t & \text{if } u_{ij}^t > q(u_{ij}, \alpha) \\ 0 & \text{if } u_{ij}^t \leq q(u_{ij}, \alpha) \end{cases} \quad t = 1, \dots, n$$

with $q(x, \alpha)$ quantile of order α for x . This formula seems to be a reasonable compromise since the threshold $q(u_{ij}, \alpha)$ varies depending on the distributive characteristics of the vector u_{ij} despite the specification of a single scalar α . On the one hand a high value for α allows only the communities characterised by remarkable weights for the sample to be detected; on the other hand it does not discriminate between connections less than or equal to $q(u_{ij}, \alpha)$. Contrariwise, a low value for α involves the critical issue of the previous example but does not involve loss of useful

information. At the interpretative level, the two proposed adjustments - normalization first, followed by introduction of threshold - relate to two jointly necessary aspects: the purpose of the former is to unify the range of observable values in order not to penalise the connections between pairs of nodes that are sparser than others due to the population's characteristics; conversely, the goal of the latter is to limit the negative drawbacks of the Louvain method due to its application *network by network*. Finally, since the methodology is aimed at creating clusters composed of networks with similar connections between nodes, the weight of every self-loop is set to zero. Indeed, two networks whose adjacency matrices share the entries outside the main diagonal could present deeply different community structures if the diagonal elements are taken into account, an aspect that inevitably causes undesirable effects in the reliability of the clustering method. Therefore the processed network R' has the same nodes as starting network R , whereas the weights are w'_{ij} instead of w_{ij} and $w'_{ii} = 0 \forall i, j, t$.

2.2 Second phase

Starting from the processed sample, it is possible to detect the n community structures by using the Louvain method. Clearly, the interpretation of the community structures changes radically when considering the new networks rather than the starting ones: any community denotes a relationship between nodes which is remarkable for the population and not necessarily identifiable by the analysis of the original networks. The tool used to define the similarity between a pair of partitions is the Adjusted Rand Index (ARI, [4]). Starting from an initial situation in which each unit is assigned to a separate group and by computing this index for each pair of partitions it is possible to create a similarity matrix and so to merge the groups using the UPGMA method [8]: the procedure ends when all the networks belong to the same cluster. The choice of the number of clusters N_G - if it is not known a priori - represents the last fundamental step: in this context, the analysis of the corresponding dendrogram, accompanied in case by qualitative evaluations by the analyst, is common practice for the hierarchical methods.

3 Results

The first aspect to investigate in the application of the clustering method to the data presented in Sect. 1 is the choice of α : since we do not have either external information or a training set for set α , it is based on the balance between the two contrasting features presented in Sect. 2.1. The selected value for α is 0.95. Such a high value is aimed at sacrificing deep connections detected in the data - all those less than or equal to the ninety-fifth percentile - in order to obtain community structures characterised by incredibly strong relationships between nodes - all those greater than the

ninety-fifth percentile -: in so doing, the problem related to the network by network application of the Louvain method is considerably reduced. On the other hand, the methodology does not take into account the obvious heterogeneity in the large set of values which are less than or equal to the threshold.

The selection of the number of groups is based on the balance of statistical criteria and practical necessities. As explained in Sect. 2.2, the analysis of the dendrogram, and in particular the evaluation of the difference between the maximum values of similarity in two successive steps of the method, is common practice when choosing N_G which, on the other hand, should not be too high to allow the analysis of the partition obtained. In the case of 26 groups, the maximum ARI is less than 0.10 (ARI=0.092), a decidedly low value: as a consequence, the selected value is $N_G = 15$ in order, at least, to obtain a manageable number of clusters. As regards the interpretation of the groups, the objective criterion used to identify the main characteristics of a specific cluster consists of the evaluation of the percentage of times that, considering the networks of that group, pairs and triads of nodes are allocated to the same community: a high value for this quantity means that the cluster is made up of playing styles with frequent connections between those pairs and triads of areas. Combining this criterion and football considerations, the 15 clusters can be divided into 6 categories identifying the main macro typologies of playing styles (PSs): sparse PS, long ball PS, dense PS, mixed PS, rapid attack PS and on-the-wings PS. For the sake of brevity, the description of the groups is not reported here: for all the details see Appendix C of [2].

3.1 Relevance of the on-the-wings playing style

The clustering method can be used, for example, to evaluate the efficacy of a specific playing style. The purpose of this section is to verify whether and how the widely used playing style with frequent connections between the lateral zones of the defense and the adjacent lateral zones of the midfield (*i.e.* the abovementioned on-the-wings playing style) impacts on the number of goals scored. In order to verify this evidence, we develop an extension of the Dixon and Coles model [3], to which we cross-refer for all the details. Let $X_k \sim \text{Poisson}(\lambda_k)$ ($Y_k \sim \text{Poisson}(\mu_k)$) be the number of goals scored by the home team (away team) in match k with $k = 1, \dots, 380$. By adding the information about the on-the-wings playing style to the formulation used by Dixon and Coles, we define:

$$\log(\lambda_k) = \gamma + \alpha_{i(k)} + \beta_{j(k)} + \delta c_{i(k)} \quad \log(\mu_k) = \alpha_{j(k)} + \beta_{i(k)} + \delta c_{j(k)}$$

with γ parameter which allows for the home effect, $\alpha_{i(k)}, \beta_{i(k)}$ parameters which measure respectively the attack and defence rates of the teams, $i(k), j(k)$ indices which identify the home and away teams, $c_{i(k)} = 1$ if the playing style of team i during match k is the on-the-wings playing style, 0 otherwise. Since the quantity of interest is δ , *i.e.* the effect of the on-the-wings playing style on the number of goals

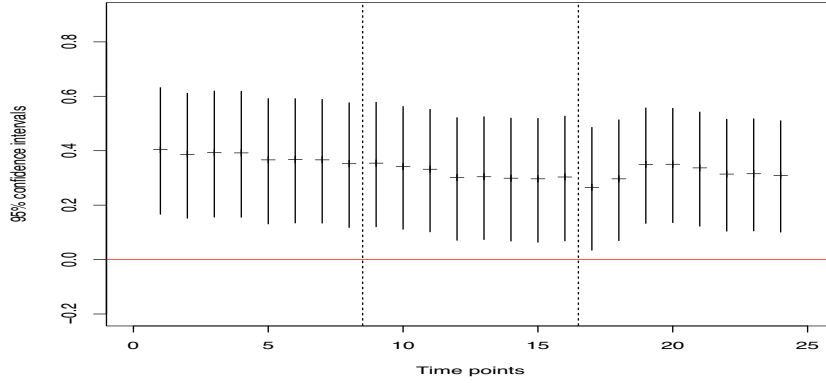


Fig. 3 95% confidence intervals and point estimates (horizontal dashes) for parameter δ . The dashed lines divide the time points into months (February-March-April).

scored, the inference focuses only on this parameter. Starting from the ‘pseudolikelihood’ $L_t(\alpha_i, \beta_i, \rho, \gamma; i = 1, \dots, 20)$ used by Dixon and Coles for each time point t and by adding the new parameter δ , we can define the following composite profile likelihood (CPL):

$$L_t^P(\delta) = L_t(\delta, \hat{\alpha}_{i_\delta}, \hat{\beta}_{i_\delta}, \hat{\rho}_\delta, \hat{\gamma}_\delta) = L_t(\delta, \hat{\theta}_\delta)$$

with ρ parameter used by Dixon and Coles to manage the dependence between X_k and Y_k and θ set of nuisance parameters. In so doing, it is possible to calculate confidence intervals based on the composite likelihood ratio statistic [9]. Let us define $\psi = (\delta, \theta)$, $u_t(\psi) = \nabla_\psi \log L_t(\psi) = \nabla_\psi l_t(\psi)$, the sensitivity matrix at time t as $H_t(\psi) = E_\psi \{-\nabla_\psi u_t(\psi)\}$, the variability matrix at time t as $J_t(\psi) = \text{var}_\psi \{u_t(\psi)\}$, the Godambe information matrix at time t as $G_t(\psi) = H_t(\psi)J_t(\psi)^{-1}H_t(\psi)$ and with $H_t^{\delta\delta}$ ($J_t^{\delta\delta}$, $G_t^{\delta\delta}$ respectively) the inverse of $H_t(\psi)$ ($J_t(\psi)$, $G_t(\psi)$ respectively) pertaining to δ . Hence, according to the asymptotic result proposed by [7], the confidence intervals are defined as follows:

$$\left\{ \delta : \frac{2(l_t^P(\hat{\delta}) - l_t^P(\delta))}{(H_t^{\delta\delta})^{-1}G_t^{\delta\delta}} < \chi_{1;1-\alpha}^2 \right\}$$

The estimates of $H_t(\psi)$ and $J_t(\psi)$ are computed according to the procedure proposed by [9] when the sample size is large. The confidence intervals are calculated from February 14th, 2016 to April 17th, 2016 for an overall amount of 24 distinct matchdays, while the start of the season is used to optimize the choice of the parameter ξ (a parameter which allows the changes in team performance during the season to be included, see [3]) through the procedure described by Dixon and Coles. Figure 3 shows the 95% confidence intervals. All the intervals, although they should be approached with particular caution since the CPLs are not independent, include

only positive values: the on-the-wings playing style seems to affect the number of goals scored positively. While maintaining its typical oscillatory trend, the sequence of point estimates is a decreasing one: a possible reason is that after an initial adjustment period, the teams take technical-tactical precautions to curb this tactic, inducing a decrease in its attacking effectiveness.

4 Conclusion

We have developed a hierarchical clustering method to divide a sample of undirected weighted networks into groups. The methodology represents a flexible solution since it allows the characteristics of the population to be managed on a case-by-case basis, and so a broad range of situations can be properly addressed. Conversely, the balance of the trade-off between the loss of information and the reliability of community structures is an undeniable limit in an unsupervised context. The application of the procedure to the analysis of playing styles provides encouraging results: the building up of the offensive manoeuvre from the lateral zones of the field seems to positively affect the number of goals scored by a football team. The high internal heterogeneity of groups - necessary to obtain a manageable number of clusters - obviously lessens the effectiveness of the approach, and consequently the effect of the tactics detected on the final outcome of a match.

Acknowledgements I am grateful to my master's thesis supervisor Bruno Scarpa for his great support, and to Lazar Petrov, Italian lead of InStat, and Lorenzo Favaro, CEO of SportAnalisi, who provided data. I also thank Daniele Durante and Nicola Sartori for fundamental comments, and David Dandolo for providing the code on the Dixon and Coles model.

References

1. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**(10), P10008 (2008)
2. Diquigiovanni, J., Scarpa, B.: Analysis of association football playing styles: An innovative method to cluster networks. *Stat. Model.* **19**(1), 28–54 (2019)
3. Dixon, M.J., Coles, S.G.: Modelling association football scores and inefficiencies in the football betting market. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **46**(2), 265–280 (1997)
4. Hubert, L., Arabie, P.: Comparing partitions. *J. Classification.* **2**(1), 193–218 (1985)
5. Newman, M.: Matrix algorithms and graph partitioning. In Newman, M.: *Networks: an introduction*, pages 345–394. Oxford Univ. Press (2010)
6. Peña, J.L., Touchette, H.: A network theory analysis of football strategies. *arXiv:1206.6904* (2012)
7. Satterthwaite, F. E.: An approximate distribution of estimates of variance components. *Biometrics bulletin.* **2**(6), 110–114 (1946)
8. Sokal, R. R., Michener, C. D.: A statistical method for evaluating systematic relationship. *Univ. Kans. sci. bull.* **38**, 1409–1438 (1958)
9. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. *Statist. Sinica.* **21**(1), 5–42 (2011)

Big data in longitudinal observational studies: how to deal with non-probability samples and technological changes

I Big data negli studi longitudinali: come trattare campioni non probabilistici e cambi di tecnologia

Clelia Di Serio, Luca Del Core, Eugenio Montini and Andrea Calabria

Abstract Big data in biomedicine represents an important example of non-probability sample. Thus, it is critical to adjust for the selection bias in analyzing big data (like Electronic Medical Records) for finite population inference both in basic as well as in clinical research. A crucial problem related to poor representativeness of target population arises when longitudinal outcomes are observed for a long period of time. Indeed, several sources of bias may arise also due to technological changes. Unreliability or under-representativeness in data may be due to machine/software or human variances/errors, or other unidentifiable external factors. Thus, two problems should be faced: (1) identify and filter-out unreliable data, and (2) harmonize samples gauged with different platforms improved over time. This work is aimed at developing a new statistical framework based on resampling techniques, such as rarefaction methods, to address both issues, showing results in real case scenarios.

Abstract *I Big Data in biomedicina rappresentano un importante esempio di campione non probabilistico. Nell'analisi dei big data (per esempio, i dati delle cartelle cliniche elettroniche) è fondamentale tener conto di eventuali bias di selezione per poter fare inferenza da popolazioni finite, sia nell'abito della ricerca di base che in quella clinica. Un problema cruciale legato alla scarsa rappresentatività della popolazione target si pone quando l'outcome di interesse è rilevato e osservato per*

Clelia Di Serio

University Vita-Salute San Raffaele, University Centre of Statistics in the Biomedical Science, Milan, Italy e-mail: diserio.clelia@univr.it

Luca Del Core

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands e-mail: l.del.core@rug.nl

Eugenio Montini

San Raffaele Telethon Institute for Gene Therapy (SR-Tiget), IRCCS San Raffaele Scientific Institute, Milan, Italy e-mail: montini.eugenio@hsr.it

Andrea Calabria

San Raffaele Telethon Institute for Gene Therapy (SR-Tiget), IRCCS San Raffaele Scientific Institute, Milan, Italy e-mail: calabria.andrea@hsr.it

un lungo periodo di tempo durante il quale ci possono essere dei cambi importanti nella tecnologia di rilevazione dei dati. L'inaffidabilità o la scarsa rappresentatività dei dati possono essere dovute anche all'errore/variabilità operatore-dipendente o ad altri fattori esterni non identificabili. Pertanto si devono risolvere due problemi: (1) identificare e filtrare i dati inaffidabili e (2) armonizzare i campioni ottenuti con piattaforme diverse soggette a progresso/obsolescenza tecnologica nel tempo. In questo lavoro verrà inquadrato il problema e verranno presentate nuove tecniche di ricampionamento, come i metodi di rarefazione, per risolvere entrambi i problemi, mostrando delle applicazioni su dati reali.

Key words: Big data, Gene therapy, Non-probability sampling, Rarefaction methods

1 Introduction

In analyzing big data for finite population inference, it is critical to consider that whenever data are retrieved from large medical records the bias derived from the non-random nature of the sample may be reinforced. Indeed, this leads to loose one of the basic task in research induced by designed sampling so that properties learnt by statistical analysis of measurements in the sample may be generalized to the population. Probability sampling can be used to construct valid statistical inferences for finite population parameters. Despite the merits of probability samples, Baker et al (2013) argue that when dealing with big data it becomes common to get non-probability samples, which may not represent the target population properly. The increasing prevalence of non-probability samples, such as web panels, makes methods for non-probability samples even more important. Keiding and Louis (2016) address the challenges in using non-probability samples for making inferences. Elliott and Valliant (2017) review the weighting methods for reducing the selection bias in non-probability samples. In EMR relying on a large observational studies leads unadjusted estimates to be very sensitive to confounding bias due to the different prognosis of individuals. In Gigliarano et al (2018) the effect of covariates is assessed by means of an Inverse Probability Weighting adjustment based on a Poisson regression model. However, problems related to non-representativeness of samples may arise also due to the longitudinal nature of studies involving data produced with high biological technologies. Heterogeneity may also be mainly due to technological advances which imply improvements in measurements resolution. This is the case of biomedical data, where the follow-ups of a clinical cohort of patients under treatment may last over several decades and the monitoring of patient health-care must benefit by the biotechnological improvements under continuous consolidation. In this kind of studies new statistical methods are required to understand how to distinguish whether the increased number of events is attributable to the change in technology rather than to the disease change itself. We address two types of problems:

1. understanding and separating different sources of variability, since unwanted source of variation in data may lead to high rates of false discoveries, high rates of missed discoveries, or both
2. follow-up of a clinical cohort of patients under a new treatment may last over several decades and the monitoring of patient health-care must benefit by the biotechnological advances and changes over time to better characterize the therapy evolution/outcome and supporting steering decisions on the personalized treatment

Even though several statistical methods exist for dealing with missing data or data integration with different sources (Horton and Kleinman, 2007), none of them is considering both problems in the same statistical approach. This work is proposed as a first step for addressing both problems within a re-sampling approach. In particular we relate this work to data coming from gene therapy.

2 Gene Therapy Data

Over the past decade, gene therapy (GT) has proved its potential as a next generation therapy for many diseases with unmet clinical need. GT can be exploited to overcome a cellular defect due to a mutated gene, providing a fully functional copy of it or to equip target cells with a new cellular function through genetic engineering. Most clinical approaches are based on the delivery of exogenous DNA molecules by viral vectors using, when stable gene transfer is needed, retrovirus- or lentivirus-derived systems. The advent of next generation sequencing (NGS) platforms (i.e. Roche/454 pyrosequencing and Illumina sequencing technology) substantially improved the accuracy and resolution of viral integration site (IS) analyses. The amount IS retrieved from a single experiment increased of two order of magnitude, opening novel scenarios for the exploitation of such a technique to investigate a wide spectrum of biological processes and of disease evolution. In this contribution we refer in particular to data collected over more than one decade where the effect of change of technology over time may introduce uncertainty in understanding whether the augmented genomic events (Fig. 1) represents a possible change in the disease evolution or reflects just a change in the technology. Actually, the Effects of introducing new PCR methods (DNA amplification techniques) for integration sites (IS) identification changing from LAM-PCR to SLiM-PCR lead to higher efficiency in data retrieval (>5 -20 fold), higher accuracy and precision in clonal quantification but requires clearly new methods for data harmonization and integration in aggregated measures.

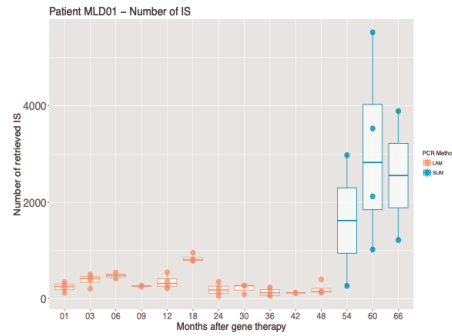
3 Methods

As mentioned two main problems should be addressed when big data coming from NGS technologies are observed for a long period: the problem of filtering, and the problem of re-sampling from a latent unique process that allows series to be comparable over time. The filtering problem is addressed through the expected richness estimation via the Hurlbert-Heck (HH) curve and the Species Pooling (SP) methods for an estimation of the unseen species. The **Rarefaction Richness** for a sample randomly (without replacement) rarefied from N to n total abundance is defined as the number of different species which we have to expect (average number):

$$E(S(n)) = S_{obs} - \sum_{i=1}^{S_{obs}} \frac{\binom{N-n_i}{n}}{\binom{N}{n}} = S_{obs} - \sum_{i=1}^{S_{obs}} q_i(n)$$

where S_{obs} represents the observed richness, n_i the abundance of i -th species and $q_i(n)$ is the probability of i -th species being not selected in the random sub-sample. The Rarefaction Curve is simply the $E(S(n))$ function defined on each value of rarefaction. The base statistical methods have been previously applied in ecological and population-based studies. We exploited a Generalized Model (GM) as an estimator of the HH curve properly rescaled. Then, using an empirical approach, we identified a minimum threshold for the richness over the whole cohort of data to filter out under-representative observations. To address the problem of data integration for reproducible results over continuous technological improvements and the scaling problem, we used a Rarefaction Method. Both methodologies have been applied in biomedical science using molecular data (retroviral vector integration sites, IS) of Gene Therapy (GT) clinical trials, a good case study for the presence of heterogeneous data. The filtering technique is used based on the richness in distinct number of ISs. The rarefaction approach allowed improving data integration of IS by rescaling data in order to obtain rarefied population measures (such as entropy indexes) that are more robust and homogeneous than the un-scaled ones, thus potentially improving the assessment of safety and long term efficacy of the treatment.

Fig. 1 Number of genomic events (integration IS) over time with the two different techniques (LAM vs SLIM).



4 Results

4.1 Filtering Unreliable Data

Dealing with biological and molecular data, such as IS in GT studies, means dealing with high variability in data collection and sampling, due to the high variability of the available biological material (for example the different amount of DNA used in each test). Thus, the number of retrieved IS from each patient at different time points (IS_s Richness) may vary. Therefore, we have to evaluate the level of richness for each sample and filter-out those samples with an insufficient level of IS richness. To overcome this problem, the percentage $S\%$ of richness in IS_s observed over the total can be defined as the ratio between the observed richness S and an estimator of the whole theoretical richness \hat{S}_{tot} , namely $S\% = S/\hat{S}_{tot}$. The theoretical richness \hat{S}_{tot} can be estimated in different ways, depending on the chosen Sampling Pooling (SP) technique (Preston, 1948; Chao, 1987; O'hara, 2005). In this work, the SP estimator is chosen based on the best Ranked Abundance Distribution (RAD) (Whittaker, 1965; Wilson, 1991) together with a p -leave out cross validation. Namely, if the general lognormal (gln) curve is chosen as the best AIC_c (Burnham and Anderson, 2004) RAD model among the candidates¹, or if, according to a χ^2 Goodness of Fit test, the gln distribution can be used in place of the optimal one, then the Preston estimator $\hat{S}_{Preston}^{tot}$ (Preston, 1948) is used. Otherwise the performance of Chao and ACE estimators is compared (Magurran and McGill, 2011) via a uniform leave- $p = .3$ -out cross validation: the whole sample is considered as the sampling universe with a known total richness S_{obs} and the $[0, 1]$ -bounded quantity defined as

$$A_{absolute}^{est} = \min\{S_{obs}, \hat{S}_{est}\} / \max\{S_{obs}, \hat{S}_{est}\}^2$$

is used to compare the performance of the two non-parametric estimators. The more the accuracy of the estimator, the greater $A_{absolute}^{est}$ is: therefore, the estimator

$$best = \arg \max_{est \in \{Chao, ACE\}} \{A_{absolute}^{est}\}$$

is chosen if the gln distribution is rejected as RAD model. Also, in order to assess the accuracy of the species pooling estimator chosen among the candidates, during the $nFld = 100$ p -leave-out cross simulations, three additional accuracy indexes defined as

¹ Geometric Series, MacArthur's Broken Stick, Zipf-Mandelbrodt, Zipf, General lognormal are the candidate RAD models in the case study. All these distributions are fitted with the Maximum Likelihood Estimation (MLE) technique.

² \hat{S}_{est} is the estimation of the total richness S_{obs} obtained using the estimator $est \in \{Chao, ACE\}$ using the 70%-random subsample.

$$A_{effective} = 1 - |.7 - S_{obs}^{0.7}/\hat{S}_{tot}^{0.7}| \quad A_{relative} = 1 - |S_{obs}^{0.7}/\hat{S}_{tot}^{0.7} - S_{obs}^1/\hat{S}_{tot}^1|$$

$$A_{cumulative} = \min \{\hat{S}_{tot}^{0.7}, \hat{S}_{tot}^1\} / \max \{\hat{S}_{tot}^{0.7}, \hat{S}_{tot}^1\}$$

are calculated³ and compared with $A_{absolute}^{best}$. By definition, they are $[0, 1]$ bounded and the more the accuracy, the greater they are. As an example, these indexes are calculated in the case study, together with $A_{absolute}^{est}$, and the results are shown in Fig.2(a). Furthermore, in order to check robustness with respect to little variations, the model selection and inference are performed on a fixed number $nRnd$ of binomial randomizations of the original data.

Therefore, a set of $S_{\%}$'s is collected among all the samples, and a minimum threshold for this quantity is identified to filter out under-representative observations. Without any ground-truth available, we used an empirical approach in which the shape of the $S_{\%}$'s empirical cumulative distribution (eCDF) is analyzed, and we selected as threshold (if existing) the main concave/convex inflection point over the eCDF curve⁴, which could be interpreted as a signal of multimodal distribution. Then, a Generalized Pareto Distribution (GPD) is fitted on the excesses above that threshold (POT method, Salvadori et al, 2007) and a QQ-plot between the empirical and GPD quantiles could provide feedbacks on the quality of the fitting such that an higher quality corresponds to a better overlap of the curve to the main diagonal. This method was applied in our case study and the results are shown in Fig.2(b,c). Another interesting question to deal with is the saturation problem: finding the total abundance Ab_{tot} needed to reach a certain percentage level p of richness, say the 90%. For this reason, the Hurlbert-Heck (HH) rarefaction curve (Hurlbert, 1971; Heck Jr et al, 1975) $E(S)$ is calculated, over a properly chosen grid of rarefaction levels of total abundance, as a pointwise estimation of the expected richness associated with each rarefaction level. Then a family of generalized models defined by

$$GM : \begin{cases} E(Y) = g^{-1}(\alpha \log(X) + \beta X) & Y \sim Binomial \\ g \in \{probit, logit, cloglog\} \\ Y = S^{\%} = E(S)/\hat{S}_{tot} & X = Ab_{tot} \end{cases}$$

are applied and the one with the maximum R^2 index is chosen. Also, in order to calculate the total abundance $Ab_{tot}^{p\%}$ associated with a certain percentage level p of richness, the regression function was inverted via the numerical resolution of the Lambert function defined by $W(z)e^{W(z)}, z \in \mathbb{C}$. In Fig.2(d) a graphical representation of the percentage HH curve and its regression estimator are shown for one sample of the case study.

³ $S_{obs}^{0.7}, \hat{S}_{tot}^{0.7}, S_{obs}^1, \hat{S}_{tot}^1$ are the observed and estimated (using the chosen estimator) total richness in the 70% fold and in the whole sample respectively.

⁴ The eCDF inflection points are found via the Extremum Distance Estimator (EDE) algorithm (Christopoulos, 2012).

4.2 Scaling of the Heterogeneous Data

Another problem regarding the IS_s data being analyzed in the case study concerns the reliability of data interpretation due to the different orders of magnitude in total abundance Ab_{tot} , and indeed in richness S_{obs} , of IS_s reached in each sample mainly due to the variation in resolution of the gauge instruments adopted during time.

Therefore, in order to compare any measure of safety obtained in each sample (e.g. an entropy index), the whole cohort of data should be first rescaled to the same magnitude level of total abundance. In this work, the minimum total abundance level among the samples defined as $Ab_{raremax} = \min_{samples} Ab_{tot}$ is used as rarefaction level.

Then, a rarefaction technique (Heck Jr et al, 1975), which essentially consists in random subsampling without replacement with proportional abundances defined as $p_i = Ab_i / Ab_{tot}$ ⁵ is applied in order to generate a rarefied version of that sample. This results in a more homogeneous pool of samples which can be used for entropy measures comparison during time of therapy. As an example, the Renyi Entropy Spectre (RES) (Principe, 2010) is calculated on the IS_s data of the case study, before and after rarefaction. The results are shown in Fig.2(e,f).

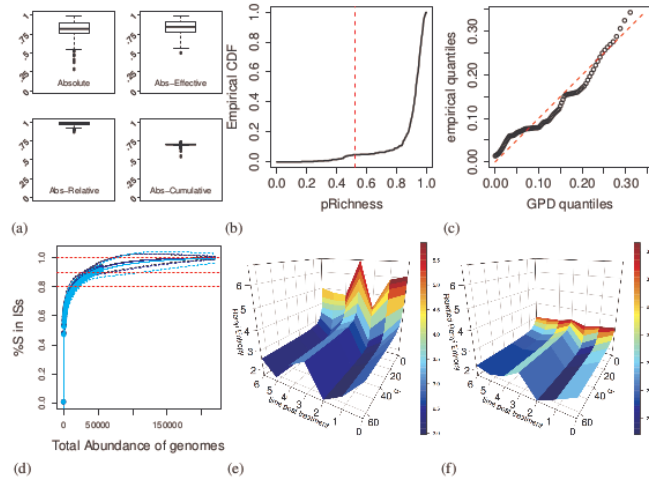


Fig. 2 (a) From top-left, the boxplot of the Absolute, Effective, Relative and Cumulative accuracy indexes are shown respectively. (b) The empirical CDF associated with the collected sample of $S^{\%}$. The vertical red highlighted line represents the threshold $S^{\%}_{thr}$ estimated via the EDE algorithm. (c) The scatterplot between the empirical quantiles associated with the excesses below that threshold and the GPD quantiles fitted on the same quantities. (d) The Hurlbert-Heck curve rescaled by the sampling pooling species estimators $S^{tot}_{Preston}$, S^{tot}_{Chao} , S^{tot}_{ACE} are drawn in black, blue and cyan. The observed and predicted ratio $S^{\%} = S_{obs}/S^{tot}$ are respectively represented by the thick and thin lines. The 80%, 90% and 100% richness thresholds are also shown as red dotted lines. (e) The Renyi Entropy Spectre is shown during time of therapy on the heterogeneous (un-rarefied) and (f) homogeneous (rarefied) samples respectively.

⁵ Ab_i is the abundance of the i -th species in the original sample.

References

- Baker R, Brick JM, Bates NA, Battaglia M, Couper MP, Dever JA, Gile K, Tourangeau R (2013) Summary report of the aapor task force on non-probability sampling. *Journal of Survey Statistics and Methodology* 1(2):90–143
- Burnham KP, Anderson DR (2004) Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research* 33(2):261–304
- Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 43(4):783–791
- Christopoulos DT (2012) Developing methods for identifying the inflection point of a convex/concave curve. *arXiv preprint arXiv:12065478*
- Elliott MR, Valliant R (2017) Inference for nonprobability samples. *Statistical Science* 32(2):249–264
- Gigliarano C, Nonis A, Briganti A, Bonetti M, Di Serio C (2018) Effect of the number of removed lymph nodes on prostate cancer recurrence and survival: evidence from an observational study. *BMC bioinformatics* 19(7):200
- Heck Jr KL, van Belle G, Simberloff D (1975) Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* 56(6):1459–1461
- Horton NJ, Kleinman KP (2007) Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* 61(21):79–90
- Hurlbert SH (1971) The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52(4):577–586
- Keiding N, Louis TA (2016) Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179(2):319–376
- Magurran AE, McGill BJ (2011) *Biological diversity: frontiers in measurement and assessment*. Oxford University Press
- O'hara R (2005) Species richness estimators: how many species can dance on the head of a pin? *Journal of Animal Ecology* 74(2):375–386
- Preston FW (1948) The commonness, and rarity, of species. *Ecology* 29(3):254–283
- Principe JC (2010) *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media
- Salvadori G, De Michele C, Kottegoda NT, Rosso R (2007) *Extremes in nature: an approach using copulas*, vol 56. Springer Science & Business Media
- Whittaker RH (1965) Dominance and diversity in land plant communities: numerical relations of species express the importance of competition in community function and evolution. *Science* 147(3655):250–260
- Wilson JB (1991) Methods for fitting dominance/diversity curves. *Journal of Vegetation Science* 2(1):35–46

Smart Data For Smart Health

Smart Data For Smart Health

Clelia Di Serio, Ernst C. Wit, Elena Bottinelli and Roberto Buccione

Health informatics, rapid evolution in genomics, and pervasive computing has changed many paradigm in biomedicine creating the possibility for large-scale collection, sharing, aggregation and analysis of volumes of data: this phenomenon is commonly known as big data. Data can be considered “big” either because they are obtained from collection of several medical database, such as Electronic Medical Records (EMR), or because they are revealed with new technologies, such as Next Generation Sequencing technologies, that produce high dimensional data. This leads to the crucial question: how to translate “big data” in “big information”? Issues like rapid evolution of technologies or low representativeness of sample from big surveys make very often results not comparable with other studies. Literally reproducibility is the ability of an entire experiment or study to be replicated. The definition looks straightforward but it carries lots of implications. It is not easy nor automatic to integrate information provided by big data in biomedicine and to translate these into a better knowledge of the clinical outcome. This contribution is organized as a “round table” where the keyword for translating big data in “smart” data are: data integration, for example using a network approach, and data reproducibility. *L’informatica sanitaria, l’evoluzione rapida in campo genomico e la diffusione di tecniche computazionali hanno cambiato molti paradigmi nella biomedicina consentendo di raccogliere dati su larga scala, di dividerli, di aggregarli e di analizzarli: questo fenomeno è comunemente noto come big data. I dati possono essere considerati “big” sia perché sono ottenuti come combinazione di diversi database clinici contenenti informazioni sullo stato di salute dei pazienti, come le cartelle cliniche elettroniche (Electronic Medical Record, EMR), sia perché tali dati sono derivati utilizzando nuove tecnologie, come le tecnologie di Next Generation Sequencing, che producono dati ad alta dimensionalità. Tutto questo porta a chiedersi come si possono tradurre i “big data” in “big information”. Le problematiche rela-*

Clelia Di Serio

Vita-Salute San Raffaele University, Milan e-mail: diserio.clelia@hsr.it

Prof. Ernst C. Wit

Università della Svizzera italiana, Lugano, Svizzera e-mail: wite@usi.ch

Ing. Elena Bottinelli

Chief Executive Officer of San Donato Group

Dr Roberto Buccione, Head Research Integrity Office IRCCS San Raffaele Hospital e-mail: buccione.roberto@hsr.it

tive alla rapida evoluzione delle tecnologie o alla scarsa rappresentatività del campione nelle survey rendono molto spesso i risultati ottenuti nei diversi studi non confrontabili tra loro. La riproducibilità in senso stretto è la capacità di riprodurre un intero esperimento o studio. Seppur di semplice definizione, la riproducibilità ha diverse implicazioni. Non è facile, e nemmeno automatico, integrare le informazioni derivanti dai big data in biomedicina e tradurle immediatamente in una migliore conoscenza dell'esito clinico.

Questo contributo è organizzato come una tavola rotonda dove gli aspetti cruciali per tradurre i “big data” in “smart data” sono: integrazione di diverse fonti di dati, usando un approccio basato su network, e riproducibilità dei dati

Key words: Big data, Information, Reproducibility, Sampling issues

Statistical network science in biomedical application

Speaker: Prof. Ernst C. Wit, Università della Svizzera italiana, Lugano, Svizzera

Abstract High-throughput genomic data in the field of genomics has revolutionized the research field. It is able to penetrate deeply into the functional, evolutionary and even social environment of the species under study. Often, however, existing data analytic methods are unable to capture the intricate nature of these processes. We present how statistical network science is able to capture the functional, evolutionary and social interactions of organisms through high-throughput data. We present three examples. First, recombinant Inbred Lines (RILs) derived from divergent parental lines can display extensive segregation distortion and long-range linkage disequilibrium (LD) between distant loci on same or different chromosomes. These genomic signatures are consistent with epistatic selection having acted on entire networks of interacting parental alleles during inbreeding. The reconstruction of these interaction networks from observations of pair-wise marker-marker correlations or pair-wise genotype frequency distortions is challenging as multiple testing approaches are under-powered and true long-range LD is difficult to distinguish from drift, particularly in small RIL panels. A side product of the previous study showed that the same method also clearly detected a genetic map signature. Especially for challenging polyploid organisms we have used this method to detect the genetic map. Apparently unrelated, but methodologically very similar is in what way microorganisms live in symbiotic relationships with their environment as they play a central role in many biological processes. They form a complex system of interacting species. Until now correlation-based network analysis and graphical modelling have been used to identify the putative interaction networks formed by the species of microorganisms, but these methods do not take into account all features of microbiota data. The methods have been implemented in two R packages (netgwas and rMAGMA), available to the practitioner. This is collaborative work with Pariya Behrouzi and Arnaud Cougoul.

Data-integration in improving hospitals efficacy indicators: a stakeholder perspective

Speaker: Ing. Elena Bottinelli, Chief Executive Officer of San Donato Group

Abstract Sharing and integrating data between hospitals to identify unwarranted variation, and then working with those practices to ensure that best practice is transferred, seems a logical way to improve the care we deliver. Healthcare stakeholders now have access to promising new threads of knowledge. This information is a form of “big data”, so called not only for its sheer volume but for its complexity, diversity, and timeliness. For a big healthcare group like Gruppo San Donato, which is the first healthcare group in the country, and a pioneer in multiple research fields, EMR (Electronic Medical Records) could be accessible by a politics of data-sharing which must be a priority to address problems related to variability in healthcare quality and escalating healthcare spend. For instance, researchers can mine the data to see what treatments are most effective for particular conditions, identify patterns related to drug side effects or hospital readmissions, and gain other important information that can help patients and reduce costs. Fortunately, recent technologic advances in the industry have improved their ability to work with such data, even though the files are enormous and often have different database structures and technical characteristics. These developments allow stakeholders access to a broader range of information not only on clinical data but also on activities and cost data or on patient behavior and sentiment data that describe patient activities and preferences, thus leading to a more effective and efficient management in healthcare systems in a global health perspective.

Data sharing for reproducibility in biomedical and clinical research

Speaker: Dr Roberto Buccione, Head Research Integrity Office IRCCS San Raffaele Hospital

Abstract Clinical research data sharing has an amazing potential to strengthen and accelerate academic research, the integrity of the clinical trial system and indeed, clinical practice. Data sharing is becoming a condition that must be met for publication in major basic and translational research journals and progress is now being made to extend such measures to medical journals as well. Some benefits for clinical research are obvious: researcher access to the data underlying trial results can enhance replication and reproducibility by encouraging independent confirmation and further analyses. Additional benefits include increasing academic rigor, integrity and transparency, and accelerating medical research. There are also other, perhaps less visible, advantages that include enhanced awareness of conflicts of interest in a clinical trial system in which external sponsorship is rather common. Last but not least, performing research in people implies a social contract, which includes making source data available for examination and re-use. Although there are many policy, privacy, and practical issues that must be addressed in to make clinical re-

search data sharing feasible and useful for the research community, the challenge must be met.

Detecting and classifying moments in basketball matches using sensor tracked data

Una procedura per identificare e classificare momenti di gioco in pallacanestro con l'uso di dati sensori.

Tullio Facchinetti and Rodolfo Metulini and Paola Zuccolotto

Abstract Data analytics in sports is crucial to evaluate the performance of single players and the whole team. The literature proposes a number of tools for both offence and defence scenarios. Data coming from tracking location of players, in this respect, may be used to enrich the amount of useful information. In basketball, however, actions are interleaved with inactive periods. This paper describes a methodological approach to automatically identify active periods during a game and to classify them as offensive or defensive. The method is based on the application of thresholds to players kinematic parameters, whose values undergo a “tuning” strategy similar to Receiver Operating Characteristic curves, using a “ground truth” extracted from the video of the games.

Abstract La “data analytics” è cruciale per valutare le prestazioni di singoli giocatori e dei team nello sport. La letteratura accademica ha sviluppato una serie di strumenti per le situazioni di attacco e di difesa. I dati che rilevano la posizione dei giocatori possono essere utilizzati per arricchire la quantità di informazioni utili. Nel basket, tuttavia, le azioni sono intervallate da periodi inattivi. In questo lavoro si propone un metodo per identificare i periodi attivi e classificarli come offensivi o difensivi. Il metodo si basa sull'applicazione di soglie sui parametri cinematici dei giocatori, i cui valori vengono sottoposti ad una strategia di “tuning” simile alle curve ROC in cui la “ground truth” viene estratta da un'analisi video.

Key words: Sport Analytics; GPS; Trajectories; Basketball

Tullio Facchinetti

Department of Industrial, Computer and Biomedical Engineering, University of Pavia, Via Ferrata, 1, 27100 Pavia, e-mail: tullio.facchinetti@unipv.it

Rodolfo Metulini

Department of Economic and Management, University of Brescia, Contrada Santa Chiara, 50, 25122 Brescia, e-mail: rodolfo.metulini@unibs.it

Paola Zuccolotto

Department of Economic and Management, University of Brescia, Contrada Santa Chiara, 50, 25122 Brescia, e-mail: paola.zuccolotto@unibs.it

1 Introduction

Recent years registered the rise of data analytics applied to sports. Experts in data science are nowadays employed by teams to improve strategical decisions. These strategies, related to the objective of winning the games, may regards either single players or the whole team and must consider both offensive and defensive performances. In basketball, Oliver [1] outlines most of the tools used to evaluate performance. Offensive performance can be measured in terms of shots (Zuccolotto et al. [2]) or considering other aspects of playing, such as the number of possessions per game (Kubatko et al. [3]). Sampaio et al. [4] and Paulauskas et al. [5] applied a descriptive discriminant analysis to different competitions to identify which variables best predict differences in the playing style. From a defensive perspective, Franks et al. [6] and Goldsberry and Weiss [7] introduced a new suite of defensive metrics suggesting to integrate spatial approaches and player tracking data.

The positioning and the velocity of players is an essential aspect to be considered when analysing both offensive and defensive performance. The robust statistical apparatus of National Basketball Association (NBA) supported by private companies made analysis of offensive and defensive moments relatively easy. Wu and Bornn [8] provide a tool for the visual analysis of offensive actions using a sensor data technology. Miller and Bornn [9] use the same data to catalogue NBA league strategies according to players' movements. Ball circulation during offense actions has been analysed by D'Amour et al. [10] to show that the more open shots opportunities can be generated with more frequent and faster movements of the ball. Less attention was paid to European leagues, mainly depending on the restriction on data collection, which is rarely granted to authorized operators. Metulini et al. [11, 12] and Metulini [13] use tracked data from Italian professional basketball games collected by mean of an accelerometer device and they split games into clusters of homogeneous spatial distances among players, looking for those with better team shooting performance. Metulini et al. [14], using the same data, measured the relation of surface area occupied by players in offence and in defence with the number of scored points by the team.

However, accelerometer devices track players' movements along the full game, without distinguishing between active/inactive periods and offensive/defensive possessions. A possible solution is to instruct a person to track these information during the game. However, this option can be unpractical either due to organizational issues and cost impact. For these reasons, we propose a procedure that identifies and removes inactive periods and classifies them as either offensive or defensive. Such a procedure allows a better usage of tracked data from localization systems for the aim of a better team performance analysis at support to game decision in basketball, from professional to amateur and youth leagues.

In this paper we discuss the procedure proposed by Metulini [15] and we introduce a validation strategy based on the use of a "ground truth" extracted from a video-based annotation of a sample of games.

2 Data

The tracking system collects the position and the velocity of every player during the full game length, including those waiting on the bench, along the x -axis (court length) and the y -axis (court width). The measured positions are expressed in centimeters (cm); the estimated accuracy of the tracking system is around 30 cm. Each measurement is marked by its time instant t . The tracking system is able to capture measurements at a sampling frequency of 50 Hz, corresponding to a measurement every 20 milliseconds (ms). We call the ordered set of measurements \mathbf{X} . The measurement made at the time instant t , denoted with x_t , thus contains the following information:

- The vector of the position for the i -th player along the x - and the y - axis, denoted as $pos_i(t) = \{pos_i^x(t), pos_i^y(t)\}$, measured in cm, where superscript x and y are used, respectively, for court length and court width;
- The vector of the velocity for the i -th player along the x - and the y - axis, denoted as $vel_i(t) = \{vel_i^x(t), vel_i^y(t)\}$, measured in kilometres per hour (km/h);
- The velocity for the i -th player in the court at time t , computed as $v_i(t) = \sqrt{vel_i^x(t)^2 + vel_i^y(t)^2}$.

3 The procedure

The procedure aims at removing specific measurements from \mathbf{X} according to three different criteria and to separate the game into offensive and defensive possessions. The filtering and labelling scheme is based on defining kinematic parameters related to players' positions and velocities on the International Basketball Federation (FIBA) court (Figure 1). The outcome is a reduced set of measurements denoted as \mathbf{Xr} that includes two features about the type of possession ($poss = \{\text{offensive, defensive, transition}\}$) and the ordered number of possession ($ord = \{1, 2, \dots, n\}$), respectively. In detail:

1. According to criterion 1-A, the procedure drops from \mathbf{X} all the measurements belonging to the time instant t in which the number of players inside the court is different from 5.
2. Criterion 1-B drops from \mathbf{X} the measurements when at least one player is on the free throw shooting area (FTSA) for at least a specified interval of time T_{ft} . Player i lies in the FTSA at time t if the vector $pos_i(t)$ lies in the circle C_r centred on the center of the FTSA of radius $r = 1.80m$.
3. The criterion 1-C removes those measurements where the speed of all the five players in the court is below a given threshold V_{min} , for an interval of time equal or larger than T_{vel} .
4. Criterion 2-A assigns the value of the variable $poss$ to each measurement of \mathbf{Xr} . The procedure generate the average x coordinate of the five players on the court

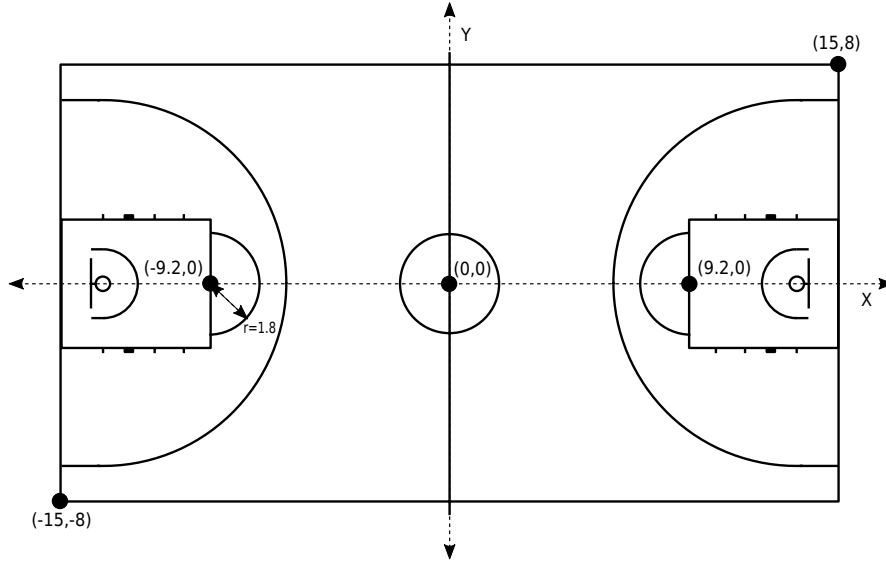


Fig. 1 International Basketball Federation (FIBA) court with relevant measures annotated.

at time t , $avg_pos^x(t) = \sum_{i=1}^5 pos_i^x(t)/5$. The measurement t could lie either on the offensive ($avg_pos^x(t) > 4$) or on the defensive ($avg_pos^x(t) < -4$) side of the court. *transition* is instead assigned to variable *poss* whereas $avg_pos^x(t)$ lies in the interval $\{+4, -4\}$.

5. Criterion 2-B assigns to the measurement of \mathbf{Xr} the *ord* value, for the aim of counting for the total number of possessions in the game. The procedure assign to *ord* value “1” to measurement x_1 . Such a value increases by 1 whenever the variable *poss* takes the value *transition* at time $t - 1$ and either *offensive* or *defensive* at time t .

4 The validation strategy

While parameter T_{ft} can easily be determined by looking to the average time required by a player to shot one or two free-throws, the “best” values for the parameters V_{min} and T_{vel} need for a tuning strategy in order to be correctly identified. The tuning strategy has the objective to find those values for V_{min} and T_{vel} such that the accordance of the procedure with the “ground truth” is maximized.

4.1 Video-based annotation

We extract the “ground truth” by using a specific smartphone application while watching the available video footage of the game. We take note of a number of game events related to i) the moments in which the game is active/inactive; ii) the moments of free-throws, time-outs, quarter- and half-time intervals; iii) the moments when the team was in offence/defence (Table 1). Based on the recorded data, we produce two reports. The first report displays when the action starts to be active

game event	description
start free-throws	a player is on the FTSA to shoot a free-throw
stop free throws	the game stops due a free-throw
start time-out	a time-out starts
stop time-out	a time-out ends
start half-time interval	an half-time interval starts
stop half-time interval	an half-time interval ends
start quarter-time interval	a quarter-time interval starts
stop quarter-time interval	a quarter-time interval ends
stop	the game stops for a generic reason
start	the game starts / restarts after a generic stop
offence	the team starts an offensive action
defence	the team starts a defensive action

Table 1 Names of the events and description.

(*action = play*) or starts to be inactive (*action = stop*) with reference to a given moment (*sec*). *active* is a variable that assumes value equal to 1 if the game starts to be inactive in that moment. *timeout*, *ft*, *quarter* and *half* are variables assuming value equal to 1 if the reason of the inactivity is, respectively, a time-out, the shooting of a free-throw, a quarter-time interval or an half-time interval. In the excerpt reported in Table 2, the game starts at second 1 (*active* = 0 & *sec* = 1 in the first row). From second 1 to second 4 the game is active. At second 5 the game stops (*active* = 1 at the second row of the table) due to a generic reason. At second 13 the game restarts (third row) and at second 47 the game stops due to a free-throw (*ft* = 1 in the fourth row). The second one reports the variable *action*, which can assumes either value 1 (start a offensive action for the team) or value 0 (start a defensive action for the team). In the excerpt in Table 3 the team starts the game in defence (*sec* = 1), it starts an offensive play at second 32, it goes back to defence at second 72, an so on.

action	sec	active	timeout	ft	quarter	half
play	1	1	0	0	0	0
stop	5	0	0	0	0	0
play	13	1	0	0	0	0
stop	47	0	0	1	0	0

Table 2 An excerpt from the first report.

action	sec	off
off	1	1
def	32	0
off	72	1
def	138	0

Table 3 An excerpt from the second report.

4.2 The “ROC” method

We use the ground truth to check for the robustness of the classification of the procedure in relation to the choice of parameters V_{min} and T_{vel} . Actually, the robustness may be evaluated either in terms of how the procedure classifies active/inactive moments or in terms of how it classifies offence and defence. We borrow the approach of the Receiver Operating Characteristic (ROC) curves. The Area Under the Curve (AUC) is traditionally computed based on sensitivities and specificities to quantify the robustness of a prediction method (Zhou et al. [16], Pepe [17], Krzanowski and Hand [18]). Sensitivity measures the proportion of *true* positives, while specificity measures the proportion of *true* negatives. The ROC and AUC help in deciding the optimal threshold value by computing sensitivity and specificity on a series of possible thresholds. In this problem we have no threshold values to set for the underlying probabilities. In our case, the measurements are directly classified by the procedure as positive or negative (i.e. active/inactive; offence/defence). However, the binary classification changes as parameters V_{min} and T_{vel} change. Therefore, we measure the performance of our procedure by evaluating the AUC with respect to different values of V_{min} and T_{vel} used as thresholds.

The proposed strategy is adopted for identifying the best parameters using active/inactive classification. The same strategy could be applied, making appropriate adaptations, using offence/defence classification. Let $\tilde{\mathbf{X}}$ be the set of measurements obtained from \mathbf{X} by aggregating the observations at a frequency of 1 second. We let $Y_{\tilde{t}}$ be the variable assuming value 1 if, according to the report, the game is inactive at second \tilde{t} , 0 otherwise. Moreover, for a given V_{min} and T_{vel} combination, let $Y_{\tilde{t}}^*$ be the variable assuming value 1 in \tilde{t} if the majority of the observations corresponding to that \tilde{t} was labelled as inactive by the procedure, 0 otherwise. We define true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) accordingly and we compute sensitivity and 1 - specificity. The method is defined by the following 3 steps:

1. For a given V_{min} we compute the AUC for T_{vel} varying in a range of values. The AUC is then computed for all the V_{min} in a range of values.
2. V_{min} is selected such that AUC is maximized.
3. Adopting the Youden’s index criteria (Youden’s [19], Fluss et al. [20] and Liu [21]), for the chosen V_{min} , the value of T_{vel} is selected such that the sum of sensitivity and specificity is maximized.

5 Results

We apply the method for classifying active/inactive periods to the set of measurements of one game played by a team during the Italian Basketball Cup Final Eight 2017. The game will be indicated as case study 1 (CS1) and the corresponding set, counting for 505,291 measurements, will be denoted with \mathbf{X}_1 . We first compute the

value of the AUC corresponding to different values of V_{min} . Up to a given point, the AUC increases as V_{min} increases; beyond such a point, the AUC decreases as V_{min} increases (Figure 2). The largest value of AUC is 0.8329. According to the second step, we select the value of V_{min} corresponding to the largest AUC , which is equal to 9.25km/h. Moving to the third step, we search for the value of T_{vel} that maximize the Youden's index (Figure 3). The largest Youden's index is found for $T_{vel} = 2$.

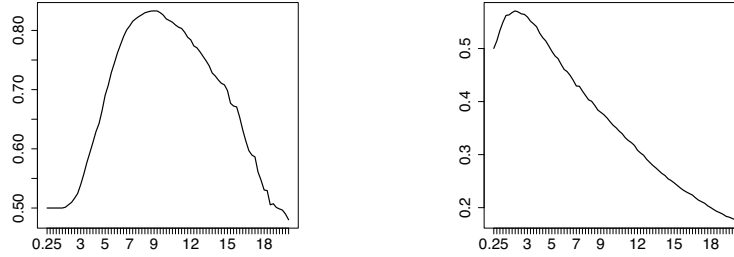


Fig. 2 AUC (y-axis) for V_{min} in $[0,20]$ (x-axis). **Fig. 3** Youden (y-axis) for T_{vel} in $[0,20]$ (x-axis).

6 Conclusions

In this work we outline a methodological approach that should be used to automatically select the correct portion of spatial tracking data of a game by just selecting those that correspond to active moments, and to correctly classify them by type of possession. The procedure, along with the identified values for the kinematic parameters may helps experts and analysts who want to analyse tracked data without watching the video of the game. Future work will focus on a more extensive application of this methodological approach and to a larger number of real case studies. First by focusing on finding the parameters that best classify active and inactive moments, then by looking to the best choice with regards to the classification among offensive and defensive possessions.

Acknowledgements Research carried out in collaboration with the Big&Open Data Innovation Laboratory (BODaI-Lab), University of Brescia (project nr. 03-2016, title *Big Data Analytics in Sports*, <https://bdsports.unibs.it/>), granted by Fondazione Cariplo and Regione Lombardia.

References

1. Oliver, D. Basketball on paper: rules and tools for performance analysis. Potomac Books, Inc. (2004).
2. Zuccolotto, P., Manisera, M., & Sandri, M. Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions. *International Journal of Sports Science & Coaching*, vol. 13(4), pp. 569-589 (2018).
3. Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, vol. 3(3) (2007)
4. Sampaio, J., McGarry, T., Calleja-Gonzalez, J., Siz, S. J., i del Alczar, X. S., & Balciunas, M. Exploring game performance in the National Basketball Association using player tracking data. *PloS one*, 10(7), e0132894. (2015)
5. Paulauskas, R., Masiulis, N., Vaquera, A., Figueira, B., & Sampaio, J. Basketball game-related statistics that discriminate between European players competing in the NBA and in the Euroleague. *Journal of human kinetics*, vol. 65, pp. 225-233 (2018).
6. Franks, A., Miller, A., Bornn, L., & Goldsberry, K. Counterpoints: Advanced defensive metrics for nba basketball. In 9th Annual MIT Sloan Sports Analytics Conference, Boston, MA. (2015)
7. Goldsberry, K., & Weiss, E. The Dwight effect: A new ensemble of interior defense analytics for the NBA. *Sports Aptitude*, LLC. Web, pp. 1-11 (2013)
8. Wu, S., & Bornn, L. Modeling offensive player movement in professional basketball. *The American Statistician*, vol. 72(1), pp. 72-79 (2018).
9. Miller, A. C., & Bornn, L. Possession sketches: Mapping nba strategies. In MIT Sloan Sports Analytics Conference (2017).
10. DAmour, A., Cervone, D., Bornn, L., & Goldsberry, K. Move or die: How ball movement creates open shots in the NBA. MIT Sloan Sports Analytics Conference (2015).
11. Metulini, R., Marisera, M. and Zuccolotto, P. Space-Time Analysis of Movements in Basketball using Sensor Data. *Statistics and data science: new challenges, new generations* (2017).
12. Metulini, R., Manisera, M. and Zuccolotto, P. Sensor Analytics in Basketball. *Proceedings of MathSport International 2017 Conference.*(2017).
13. Metulini, R. Players Movements and Team Shooting Performance: a Data Mining approach for Basketball. Book of short papers SIS2018, ISBN-9788891910233. Publisher: Pearson, Editors: Antonino Abbruzzo, Eugenio Brentari, Marcello Chiodi e Davide Piacentino (2018).
14. Metulini, R., Manisera, M., Zuccolotto, P. Modelling the dynamic pattern of surface area in basketball and its effects on team performance. *Journal of Quantitative Analysis in Sports*, vol. 14.3, pp. 117-130 (2018).
15. Metulini, R. Filtering procedures for sensor data in basketball. *Statistics & Applicazioni*, vol. 15(2), pp. 133-150 (2017).
16. Zhou, X. H., McClish, D. K., & Obuchowski, N. A. Statistical methods in diagnostic medicine (vol. 569). John Wiley & Sons (2009).
17. Pepe, M. S. The statistical evaluation of medical tests for classification and prediction. *Medicine* (2003).
18. Krzanowski, W. J., & Hand, D. J. ROC curves for continuous data. Chapman and Hall/CRC (2009).
19. Youden, W. J. Index for rating diagnostic tests. *Cancer*, vol. 3(1), pp. 32-35 (1950).
20. Fluss, R., Faraggi, D., & Reiser, B. Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 47(4), pp. 458-472 (2005).
21. Liu, X. Classification accuracy and cut point selection. *Statistics in medicine*, vol. 31(23), pp. 2676-2686 (2012).

Ordered response models for cyber risk

Modelli a risposta ordinale per la valutazione del cyber risk

Silvia Facchinetti and Claudia Tarantola

Abstract In the last years there have been a scholars increasing interest in cyber-security risk measurement, data security, and privacy protection. Since quantitative loss data are rarely available, we deal with ordinal data representing experts' evaluation of the severity of the attacks. Due to the ordinal nature of the available data, it turns natural to rely on cumulative link models that allows us to express the cumulative probabilities associated with the different severity levels as a non linear function of a suitable set of explanatory variables. We evaluate the effect of each explanatory categorical variable on the risk level using the *Average Marginal Effect*. We apply our model to a real data set that includes information on serious cyber attacks occurred worldwide in 2018.

Abstract Negli ultimi anni si è registrato un interesse crescente da parte degli studiosi riguardo il problema del cyber risk e la sua misurazione. Poichè i dati quantitativi sulle perdite sono raramente disponibili, essi sono spesso rilevati su scala ordinale e riguardano il livello di gravità degli attacchi cibernetici. Risulta pertanto naturale valutare il cyber risk mediante modelli di risposta ordinale che legano la variabile gravità a variabili esplicative spesso di natura categorica. L'effetto di tali variabili viene valutato utilizzando l'AME (*Average Marginal Effect*). Il modello viene applicato a dati reali sulla gravità degli attacchi rilevati nel mondo nel 2018.

Key words: ordered response models, cyber risk, ordinal variables, Average Marginal Effect

Silvia Facchinetti

Department of Statistical sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1 - Milano, e-mail: silvia.facchinetti@unicatt.it

Claudia Tarantola

Fintech laboratory, Department of Economics and Management, University of Pavia, Strada Nuova, 65 - Pavia e-mail: claudia.tarantola@unipv.it

1 Motivation

Cyber risk is an operational risk caused by attacks on Information and Communication Technologies systems (ICT), which are gaining increasing importance, due to the globalisation of financial activities as well as to the technological advancements. (see e.g. [4], [5], [9]). Institutions should be encourage to collect data on cyber incidents in order to use statistical approaches to estimating the capital needed to cover losses due to the occurrence of cyber attacks. Companies need to identify the key vulnerable assets that are exposed to cyber risk and to implement an integrated cyber security solution and optimal investment decisions that would automate and accelerate the threat defense [8].

The main problem in cyber risk measurement is the lack of available data (see [2]). Indeed, cyber loss data are very difficult to obtain since these data are very sensitive, and it is uncommon for an institution to be willing to disclose them, since it wants to preserve its reputation and security. This lack of disclosure drastically limits the data available for the analysis of cyber risks, by regulators, research centers and by third party organisations, such as insurance companies, which may offer hedging against cyber risks.

In this context, it is quite natural to employ an ordinal measurement approach, rather than a quantitative one. Cyber risk data are often collected on an ordinal scale, in terms of severity of the attacks ordered according to the corresponding magnitude, for example critical, high and medium severity (see [6]). Due to the ordinal nature of the available data, we rely on cumulative link models that allows us to express the cumulative probabilities associated with the different severity levels as a non linear function of a suitable set of explanatory variables.

We describe the methodology and we apply our model to a real data set concerning serious cyber attacks that occurred worldwide in 2018.

2 Methodology

Let Y be the categorical variable severity with ordered levels $j = 1, 2, \dots, k$. The cumulative link model is:

$$\text{link}[P(Y \leq j)] = \alpha_j - \mathbf{x}^T \boldsymbol{\beta}, \quad (1)$$

where \mathbf{x} is a p -vector of regression variables, α_j is the cut-point, $\boldsymbol{\beta}$ is the vector of the regression parameters and link is a suitable link function (see [1] for details).

The larger the value of $\mathbf{x}^T \boldsymbol{\beta}$, the higher the probability that the response will fall in a category at the upper end of the scale. Since nonlinear link function naturally produces effects on the link scale that may be not easy to interpret, we evaluate the effect of each explanatory variable on the risk level using the so-called Marginal Effect (ME) measure, see e.g. [7]. The ME measures how a change in a specific covariate affects the response variable, holding constant the value of all the other

covariates. This measure is particularly useful because it is intuitive (the slope of the regression surface with respect to a given covariate) and can be calculated from essentially any type of covariates. The ME measure is computed differently for categorical and continuous explanatory variables. For categorical variables, they measure the discrete change. In particular, for binary data they correspond to the discrete change, and indicate how predicted probabilities change as the binary independent variable goes from 0 to 1, holding all the other variables constant, while for factor levels they indicate how the probability of being in a category changes when we move from the baseline category to the considered one. For continuous variables, they measure the instantaneous rate of change.

Among the alternative versions of the ME (see e.g. [10]), we focus on the Average Marginal Effect (AME) that is obtained by calculating the marginal effect of a specific covariate at each of the n samples of the explanatory variables, and then averaging them.

3 Empirical analysis

We consider real data collected by the researcher of the Hackmanac Project and described in the "Italian Annual report on ICT Security in Italy 2019" [3], by the Clusit association (Associazione Italiana per la Sicurezza Informatica).

The complete data set consists of 8,417 worldwide observations on serious cyber attacks in the years 2011-2018. We consider for our analysis 1,552 cyber attacks that occurred in 2018, the most recent year. According to Clusit, an attack is considered serious if it had a significant impact on the victims in terms of economic losses, damages to reputation and/or dissemination of sensitive data. We consider a model that relates the severity, a three-category ordinal response variable that measures the severity of an attack (1 = critical, 2 = high, 3 = medium severity), to the following categorical explanatory variables:

- *Attack Technique* coded in 5 categories: 0-day, Multiple Threats, Trivial Threats, SQL Injection, Unknown;
- *Continent* coded in 6 categories: Africa, America, Asia, Australia/Oceania, Europe, Multiple Continents;
- *Type of Attack* coded in 4 categories: Cybercrime, Hactivism, Espionage/Sabotage, Information Warfare;
- *Victim* coded in 19 categories: Automotive, Chemical/Medical, Banking/Finance, Critical Infrastructures, Entertainment/News, GDO/Retail, Gov-Mil-LEAs - Intelligence, Gov.Contractors/Consulting, Health, Hospitality, Multiple Targets, Online Services/Cloud, Organization/ONG, Others, Religion, Research- Education, Security Industry, Software/Hardware Vendor, Telco.

To evaluate and interpret the effects of an explanatory variable on the response probability, we now calculate the Average Marginal Effect. Since only the highest and lowest probabilities change monotonically as the explanatory variable increases,

the most extreme outcomes are often of special interest (in our context critical vs medium severity). For this reason, in Table 1 we present the results only for extreme categories.

Fitting a cumulative logit link model, we noticed that not all the categories of the considered explanatory variables present a significant relation with severity, at level $\alpha = 0.05$. Moreover, there is an inverted effect for the medium severity attacks with respect to those with critical severity. For categorical variables with more than two possible values (as the variables in our data set), the MEs show the difference in the estimated probabilities for cases in a given category relative to the baseline one.

First consider the variable *Attack Technique*, with baseline category SQL Injection. All the categories present a significant relation with the severity. In particular:

- Multiple Threats are, on average, 0.999 more likely to generate a critical attack and -0.999 less likely to generate a medium severity attack than SQL Injection;
- Trivial Threats are, on average, 1.420 more likely to generate a critical attack and -2.446 less likely to generate a medium severity attack than SQL Injection;
- Unknown attack techniques are, on average, 1.423 more likely to generate a critical attack and -2.452 less likely to generate a medium severity attack than SQL Injection;
- 0-day is, on average, 3.062 more likely to generate a critical attack and -5.275 less likely to generate a medium severity attack than SQL Injection.

Next, we consider the variable *Continent*, with baseline category Multiple Continents. Only the category Asia show a significant relation with the response variable. This continent is, on average, 0.105 more likely to generate a critical attack than Multiple Continents. For the medium severity attacks we observe an inverted effect: Asia is, on average, -0.181 less likely to generate a medium severity attack than the reference category.

With reference to the variable *Type of Attack*, the categories Espionage/Sabotage and Information warfare present a significant relation with the severity, with respect to the baseline category Cybercrime:

- Espionage/Sabotage is, on average, 0.514 more likely to generate a critical attack and -0.403 less likely to generate a medium severity attack than Cybercrime;
- Information warfare is, on average, 0.310 more likely to generate a critical attack and -0.534 less likely to generate a medium severity attack than Cybercrime.

Finally, consider the variable *Victim*, with baseline category Automotive, the significant categories are 7: Chemical/Medical, Critical Infrastructures, Gov-Mil-LEAs-Intelligence, Gov. Contractors/Consulting, Health, Organization/ONG, Telco. In particular:

- Chemical/Medical is, on average, 1.504 more likely to generate a critical attack and -2.591 less likely to generate a medium severity attack than Automotive;
- Critical Infrastructures are, on average, 0.440 more likely to generate a critical attack and -0.757 less likely to generate a medium severity attack than Automotive;

Ordered response models for cyber risk

Table 1 AME for the cumulative logit model fitted to the cyber risk data.

\$ME.1 (critical severity)		effect	std.error	z.value	p.value
Attack technique	Multiple Threats	0.999	0.000	15589.237	0.000
	Trivial Threats	1.420	0.069	20.654	0.000
	Unknown	1.423	0.069	20.524	0.000
	0-day	3.062	0.135	22.672	0.000
Continent	America	0.036	0.055	0.655	0.512
	Asia	0.105	0.021	4.920	0.000
	Europe	0.028	0.020	1.401	0.161
	Australia/Oceania	0.057	0.041	1.397	0.162
Type of attack	Africa	0.036	0.052	0.692	0.489
	Espionage/Sabot.	0.514	0.073	7.073	0.000
	Hacktivism	-0.035	0.025	-1.404	0.160
	Inf. Warfare	0.310	0.038	8.096	0.000
Victim	Banking/Finance	0.210	0.110	1.913	0.056
	Chemical/Medical	1.504	0.066	22.672	0.000
	Critical Infr.	0.440	0.072	6.141	0.000
	Entert./News	0.032	0.068	0.475	0.635
	GDO/Retail	0.044	0.073	0.594	0.553
	Gov-Mil-LE-Int.	0.311	0.066	4.678	0.000
	Gov. Contr./Cons.	0.403	0.103	3.903	0.000
	Health	0.270	0.067	4.036	0.000
	Hospitability	0.029	0.073	0.400	0.689
	Multiple targets	0.058	0.066	0.883	0.377
	Online Serv/Cloud	0.085	0.067	1.256	0.209
	Organization-ONG	0.165	0.080	2.051	0.040
	Others	0.065	0.075	0.876	0.381
	Religion	-0.056	0.124	-0.456	0.649
	Research-Educ.	0.054	0.068	0.799	0.424
	Security	0.182	0.118	1.540	0.124
	SW/HW Vendor	0.092	0.068	1.352	0.176
	Telco	0.237	0.093	2.546	0.011
\$ME.3 (medium severity)		effect	std.error	z.value	p.value
Attack technique	Multiple Threats	-0.999	0.000	-99521.457	0.000
	Trivial Threats	-2.446	0.071	-34.415	0.000
	Unknown	-2.452	0.072	-33.920	0.000
	0-day	-5.275	0.123	-42.942	0.000
Continent	America	-0.063	0.092	-0.687	0.492
	Asia	-0.181	0.036	-5.089	0.000
	Europe	-0.048	0.034	-1.409	0.159
	Australia/Oceania	-0.098	0.070	-1.400	0.162
Type of attack	Africa	-0.062	0.089	-0.693	0.488
	Espionage/Sabot.	-0.403	0.020	-20.164	0.000
	Hacktivism	0.061	0.043	1.405	0.160
	Inf. Warfare	-0.534	0.068	-7.842	0.000
Victim	Banking/Finance	-0.468	0.145	-3.221	0.001
	Chemical/Medical	-2.591	0.060	-42.942	0.000
	Critical Infr.	-0.757	0.126	-5.995	0.000
	Entert./News	-0.056	0.117	-0.475	0.635
	GDO/Retail	-0.075	0.126	-0.593	0.553
	Gov-Mil-LE-Int.	-0.536	0.114	-4.703	0.000
	Gov. Contr./Cons.	-0.694	0.179	-3.877	0.000
	Health	-0.466	0.114	-4.078	0.000
	Hospitability	-0.050	0.125	-0.399	0.690
	Multiple targets	-0.101	0.114	-0.881	0.378
	Online Serv/Cloud	-0.146	0.116	-1.255	0.210
	Organization-ONG	-0.284	0.138	-2.057	0.040
	Others	-0.113	0.129	-0.875	0.382
	Religion	0.097	0.213	0.456	0.649
	Research-Educ.	-0.094	0.117	-0.797	0.425
	Security	-0.314	0.204	-1.543	0.123
	SW/HW Vendor	-0.158	0.117	-1.351	0.177
	Telco	-0.407	0.159	-2.563	0.010

- Gov-Mil-LEAs-Intelligence are, on average, 0.311 more likely to generate a critical attack and -0.536 less likely to generate a medium severity attack than Automotive;
- Gov. Contractors/Consulting are, on average, 0.403 more likely to generate a critical attack and -0.694 less likely to generate a medium severity attack than Automotive;
- Health is, on average, 0.270 more likely to generate a critical attack and -0.466 less likely to generate a medium severity attack than Automotive;
- Organization/ONG is, on average, 0.165 more likely to generate a critical attack and -0.284 less likely to generate a medium severity attack than Automotive;
- Telco is, on average, 0.237 more likely to generate a critical attack and -0.407 less likely to generate a medium severity attack than Automotive.

4 Conclusion

We have proposed a novel model for cyber risk measurement, based on cumulative logit model that allows to express the cumulative probabilities associated to the different severity levels to a non linear function of a suitable set of explanatory variables. To analyze the effect of the explanatory variables on the severity of cyber attack we exploit the Average Marginal Effect. An application of these technique to real data show how to interpret the significant effect of the considered categorical explicative variables on the level of severity of the attacks.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation program "FIN-TECH: A Financial supervision and Technology compliance training programme" under the grant agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA).

We thanks the experts of the Hackmanac Project and Clusit for sharing the data set.

References

1. Agresti, A., and Tarantola, C. (2018). Simple Ways to Interpret Effects in Modeling Ordinal Categorical Data, *Statistica Neerlandica*, 72: 210-223.
2. Afful-Dadzie, A., and Allen, T.T. (2017). Data-Driven Cyber-Vulnerability Maintenance Policies, *Journal of Quality Technology*, 46: 234-250.
3. Antonielli, A., Bechelli, L., Bosco, F., Butti, G., et al (2019). Rapporto 2019 sulla Sicurezza ICT in Italia, Clusit.

Ordered response models for cyber risk

4. Cebula, J.J., and Young, L.R. (2010). A Taxonomy of Operational Cyber Security Risks, Technical Note CMU/SEI-2010-TN-028, Software Engineering Institute, Carnegie Mellon University, 1-34.
5. Edgar, T.W., and Manz, D.O. (2017). Research Methods for Cyber Security, Elsevier.
6. Facchinetti, S., Giudici, P., and Osmetti, S.A. (2018). How to measure cybersecurity risk. In Abbruzzo A., Brentari E., Chiodi M., Piacentino D. (eds), Book of Short Papers SIS 2018, Pearson, 1643-1646.
7. Greene, W. (2008) Econometric Analysis. Upper Saddle River, NJ: Pearson Prentice Hall, 6th edn.
8. Kolfal, B., Patterson, R.A., and Yeo, M.L. (2013). Market Impact on IT Security Spending, Decision Sciences, 44: 517-556.
9. Kopp, E., Kaffenberger, L., and Wilson, C. (2017). Cyber Risk, Market Failures, and Financial Stability, IMF Working Paper WP/17/185, 1-35.
10. Long, J.S., and Freese, J. (2014). Regression models for categorical dependent variables using Stata (3rd ed.). College Station, TX: Stata Press.

Functional data analysis-based sensitivity analysis of integrated assessment Models for climate change modelling

Analisi di sensitività basata sull'analisi di dati funzionali per modelli di valutazione integrata dei cambiamenti climatici

Matteo Fontana, Massimo Tavoni and Simone Vantini

Abstract Motivated by an application in the realm of climate change economics, we develop and prove the mathematical properties of a Global Sensitivity analysis (GSA) method for function-valued responses and, by exploiting the similarity between GSA and functional analysis of variance models, we employ a domain-selective nonparametric testing technique to assess the significance of the calculated sensitivity indices. The results confirm the qualitative intuitions of previous works, determining in a quantitative way the sparsity and heterogeneity of input effects and the weak impact of interactions between inputs. Moreover, we can appreciate the complex and non-linear time-dynamics of input effects.

Abstract *Spinti da una applicazione nell'ambito dell'economia dei cambiamenti climatici, sviluppiamo e dimostriamo le proprietà di un metodo di analisi di sensitività globale per risposte funzionali. Inoltre, sfruttando la somiglianza tra tale metodo e modelli ANOVA funzionali, utilizziamo una metodologia per l'inferenza statistica, nonparametrica e in grado di selezionare le parti significative del dominio, per testare la significatività degli indici di sensitività calcolati. I risultati confermano le intuizioni qualitative presenti in lavori precedenti, mostrando in modo quantitativo la sparsità ed eterogeneità degli effetti degli input e lo scarso impatto delle interazioni. Inoltre, la nostra metodologia permette di apprezzare la dinamica temporale, complessa e nonlineare, dell'effetto degli input.*

Matteo Fontana

Department of Management, Economics and Industrial Engineering, Politecnico di Milano e-mail: matteo.fontana@polimi.it

Massimo Tavoni

Department of Management, Economics and Industrial Engineering, Politecnico di Milano e-mail: massimo.tavoni@polimi.it

Simone Vantini

MOX - Department of Mathematics, Politecnico di Milano e-mail: simone.vantini@polimi.it

Key words: Functional Data Analysis, Functional Testing, Domain Selection, Global Sensitivity Analysis, Integrated Assessment Models, Climate Change Economics, Climate Change Modelling

1 Introduction

Climate change is by far the biggest and most imminent threat to the survival of humankind: according to the last Intergovernmental Panel for Climate Change (IPCC) [2] more decisive actions must be undertaken now, if we want to contain the average increase of the world surface temperature by 1.5 °C

A fundamental tool to understand and explore the complex dynamics that regulates this phenomenon is the use of computer models. In particular, since the very early stages of the discipline with the DICE model [4], the modelling community has oriented itself towards the use of Integrated Assessment Models, complex pieces of software that, in a nutshell, integrate a climatic and an economic model, to generate predictions about decision variables for a given amount of time (usually, the next century). The usual variable of interest in this kind of analysis is the level of CO_2 emissions: a useful proxy for changes in climate variables, such as global average temperature.

Predicting a quantity for such a long time scale is a very hard task, with a great degree of uncertainty involved. Many efforts are being undertaken to model and control this uncertainty, such as the development of standardized scenarios of future development, called Shared Socioeconomic Pathways (SSPs) [5, 8] or the use of IAM ensembles to tackle the issue of model uncertainty. Given also the relative opaqueness and the complexity of IAMs, post-hoc diagnostic methods have been used to post-process IAM outputs, for instance with the purpose of performing Global Sensitivity Analysis ([3] and references therein).

This work addresses two shortcomings in the current literature, both theoretical and applied: the first one that, even if it is very natural to interpret the output of a IAM, or, in general, of any model that yields a set of time-series or other variables indexed on a domain, and not a single point, as a function on this domain, to our knowledge there are no GSA methods able to deal with functional-valued outputs. Probably because of this absence, there are no GSA studies that focus their attention over the time dynamics of sensitivity. This fact represent a serious void in the literature, given the extreme importance of the time dimension in our analysis and understanding of climate change, and the increasing reliance on computer experiments that yield complex data as outputs in many fields of the social and physical sciences.

The second point is about the formal treatment of uncertainty in GSA: formal statistical testing procedures are rarely employed to assess and quantify model uncertainty.

2 Methods and Analysis

To address the two shortcomings identified in the introduction, we move from [3] and extend it to the proper embedding of IAM outputs: the time dimension. In a more specific sense, we treat IAM outputs as square integrable functions on $T = [2010, 2090]$, and then develop, in the spirit of [6], a unique finite decomposition framework for functional-valued outputs. We thus employ this decomposition to define absolute and normalized sensitivity indices. By exploiting the similarity between the proposed SA technique and Functional ANOVA models [7] we show that the sensitivity indices can be seen as the coefficients of a linear model with a functional response, and by using the domain-selective testing framework for linear models proposed by [1], we are able to test the indices significance. More details about the proposed methods will be available in a forthcoming MOX-Report.

We apply this analysis pipeline to the data used in [3]. We first extract the CO_2 emissions profiles on $T = [2010, 2090]$ by using a smoothing spline approach, and then, by using a small-dimensional parametrization of SSPs, we perform sensitivity analysis and testing.

3 Conclusion

The application of such method unveils novel insight inside the dynamics of CO_2 emissions and relative sensitivity measures, yielding non-linearities, non-monotonicities and, in general, an unexpected behaviour that could not be captured by standard univariate global sensitivity measures. The testing effort provides even more interesting results, showing differences between the two contrasts analyzed in this paper, and, in general, identifying a sparsity in effects: The only significant factors in determining CO_2 emissions seem to be GDP per capita and energy intensity improvements, with fossil fuel availability being significant only in the contrast between the middle-of-the-road scenario and *SSP3*. There is no statistical evidence to affirm that interaction terms are significant, with the only notable "near-miss" of the interactions that involve GDP per capita. This is probably due to the pervasiveness and centrality of GDP as the main economic variable inside economic models. This findings provide a very strong signal to the IAM community that either the Shared Socioeconomic Pathways are too refined to be actually significant inside a representative ensemble of models, or that, while preserving their own individuality and peculiarities in the modelling approach, that IAMs need to converge towards more homogeneous predictions.

Acknowledgements

Matteo Fontana and Massimo Tavoni acknowledge financial support from the European Research Council, ERC grant agreement no 336155 - project COBHAM “The role of consumer behaviour and heterogeneity in the integrated assessment of energy and climate policies”

References

- [1] Konrad Abramowicz et al. “Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament”. en. In: *Scandinavian Journal of Statistics* 45.4 (Dec. 2018), pp. 1036–1061. ISSN: 1467-9469. DOI: 10.1111/sjos.12333. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12333> (visited on 02/12/2019).
- [2] IPCC. *Global Warming of 1.5 °C. An IPCC Special Report on the impacts of global warming of 1.5 °C above pre-industrial levels and related global green-house gas emission pathways, in the context of strengthening the global re-sponse to the threat of climate change, sustainable development, and efforts to eradicate poverty*. 2018.
- [3] G. Marangoni et al. “Sensitivity of projected long-term CO₂ emissions across the Shared Socioeconomic Pathways”. In: *Nature Climate Change* 7.2 (Feb. 2017), pp. 113–117. ISSN: 1758-678X. DOI: 10.1038/nclimate3199. URL: <http://www.nature.com/articles/nclimate3199>.
- [4] William D. Nordhaus. “Rolling the ‘DICE’: an optimal transition path for controlling greenhouse gases”. In: *Resource and Energy Economics* 15.1 (Mar. 1993), pp. 27–50. ISSN: 0928-7655. DOI: 10.1016/0928-7655(93)90017-0. URL: <https://www.sciencedirect.com/science/article/pii/0928765593900170>.
- [5] Brian C. O’Neill et al. “A new scenario framework for climate change research: the concept of shared socioeconomic pathways”. In: *Climatic Change* 122.3 (Feb. 2014), pp. 387–400. ISSN: 0165-0009. DOI: 10.1007/s10584-013-0905-2. URL: <http://link.springer.com/10.1007/s10584-013-0905-2>.
- [6] Herschel Rabitz and Ömer F. Aliş. “General foundations of high-dimensional model representations”. In: *Journal of Mathematical Chemistry* 25.2/3 (1999), pp. 197–233. ISSN: 02599791. DOI: 10.1023/A:1019188517934. URL: <http://link.springer.com/10.1023/A:1019188517934>.
- [7] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer, 2005, p. 426. ISBN: 978-0-387-22751-1.
- [8] Keywan Riahi et al. “The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview”. In: *Global Environmental Change* 42 (Jan. 2017), pp. 153–168. ISSN: 0959-3780.

REFERENCES

DOI: 10.1016/J.GLOENVCHA.2016.05.009. URL: <https://www.sciencedirect.com/science/article/pii/S0959378016300681>.

Coupled Gaussian Processes for Functional Data Analysis

Processi gaussiani per l'analisi dei dati funzionali

L. Fontanella, S. Fontanella, R. Ignaccolo, L. Ippoliti, P. Valentini

Abstract We present an approach for modelling multivariate dependent functional data. To account for the dominant structural features of the data, we rely on the theory of Gaussian Processes and extend hierarchical dynamic linear models for multivariate time series to the functional data setting. We illustrate the proposed methodology within the framework of bivariate functional data and discuss problems referring to detection of spatial patterns and curve prediction.

Abstract *In questo lavoro, viene presentato un approccio idoneo alla modellazione di dati funzionali multivariati che presentano dipendenza. Per considerare le caratteristiche strutturali dominanti dei dati, ci si avvale della teoria dei processi gaussiani e si considera l'estensione dei modelli lineari dinamici gerarchici per serie storiche nell'ambito dei dati funzionali. La metodologia proposta viene illustrata con riferimento a dati funzionali bivariati. Inoltre, vengono esaminati i problemi connessi all'individuazione di patterns spaziali e previsioni funzionali.*

Key words: Gaussian Processes, Functional data, Simultaneous diagonalization, Derivative process

L. Fontanella
University G. d'Annunzio, Chieti-Pescara, Italy, e-mail: lfontan@unich.it

S. Fontanella
University of Torino, Italy e-mail: sara.fontanella@unito.it

R. Ignaccolo
University of Torino, Italy e-mail: rosaria.ignaccolo@unito.it

L. Ippoliti
University G. d'Annunzio, Chieti-Pescara, Italy e-mail: ippoliti@unich.it

P. Valentini
University G. d'Annunzio, Chieti-Pescara, Italy e-mail: pvalent@unich.it

1 Introduction

Effective statistical modelling under complex designs for functional data is still under development and requires innovative theories. In the following we focus on multivariate *dependent* functional data where the dependence can arise via multiple responses, temporal or spatial effects. In particular, we consider two different problems for bivariate functional data and illustrate the proposed methodology in the frameworks of detection of spatial patterns and curve prediction. Specific applications within these frameworks will be discussed in an extended version of the present article.

2 Detection of spatial patterns using coupled GPs

We focus on the identification of patterns of oscillations considering the *simultaneous orthogonal* expansions of Gaussian Processes (GPs). Orthogonal expansion of GPs has been extensively used for both theoretically investigation and applications. In the case of univariate processes, the theory is based on the probabilistic corollary of Mercer's theorem which is known as Karhunen-Loève expansion. Following Root [2], we consider an extension of this expansion to the case of two kernels that allows the simultaneous orthogonal expansion of two Gaussian processes. Specifically, we consider the functional data $Y(\mathbf{s}, \tau)$, where $\mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^2$ is a continuous spatial index and $\tau \in \mathcal{T} \subseteq \mathcal{Z}$ is an index of time, as a sample function of either one of two zero-mean Gaussian processes, with spatial covariances $Q_1(\mathbf{s}, \mathbf{s}')$ and $Q_2(\mathbf{s}, \mathbf{s}')$. Then, it can be shown [2] that the observed process can be expanded in terms of a series of *spatial patterns*, $w_k(\mathbf{s}) = (Q_1^{1/2} u_k)(\mathbf{s})$ as

$$Y(\mathbf{s}, \tau) = \sum_{k=1}^{\infty} \alpha_{y,k}(\tau) w_k(\mathbf{s}) \quad (1)$$

where $\alpha_{y,k}(\tau)$ is a sequence of independent Gaussian variables and u_k are orthonormalized eigenfunctions in the domain of $Q_1^{-1/2}$.

If only a limited number, K , of patterns are considered, we have a truncated expansion of equation (1), which leads to the following measurement equation

$$\begin{aligned} Y(\mathbf{s}, \tau) &= \sum_{k=1}^K \alpha_{y,k}(\tau) w_k(\mathbf{s}) + \sum_{k=K+1}^{\infty} \alpha_{y,k}(\tau) w_k(\mathbf{s}) \\ &= Y^{(K)}(\mathbf{s}, \tau) + \mathbf{e}(\mathbf{s}, \tau). \end{aligned} \quad (2)$$

Equation (2) allows for extending hierarchical dynamic linear models for multivariate time series to the functional data setting. Moreover, it can be shown that the proposed framework also allows to leverage knowledge from one process when

solving an inferential task for another and, accordingly, it forms the motivation for *transfer learning* and prediction of functional data at new sites.

3 Interpolation using derivatives

For the sake of simplicity, consider the problem of representing a one-dimensional real function $Y(\tau)$, $\tau \in \mathcal{T} \subseteq \mathcal{R}$, or estimating its derivative, $X(\tau)$, using only a limited amount of data at points, $\tau_1, \tau_2, \dots, \tau_n$. This is useful in many applications, including shape analysis, Bayesian optimization and reconstruction of surfaces and signals.

Suppose that all the known values \mathbf{y} and derivatives \mathbf{x} are collected into a vector \mathbf{z} . Let κ be a vector of corresponding indices to show the order of the derivative. For each site τ_j there may be several choices of κ_j if the value of the function and of some of its derivatives are all known at that site. For a bivariate process, $\kappa_i = 0$ if $z(\tau_i) = y(\tau_i)$ is a data value, and $\kappa_i = 1$ if $z(\tau_i) = x(\tau_i)$ is a first derivative.

Denote with $\mathbf{U} = \{u_{im}\}$ the (n, r) matrix of drift terms

$$u_{im} = \frac{\partial^{|\kappa_i|}}{\partial \tau_i^{\kappa_i}}(\tau_i^m), \quad 1 \leq i \leq n, \quad |m| \leq r,$$

where r is the order of the polynomial drift, and with u_0 the vector of drift terms at a new site τ_0 . Assume also that $\Sigma = \{\sigma_{ij}\}$ is a non-singular block covariance matrix where each block has entries given by

$$\sigma_{ij} = (-1)^{|\kappa_j|} \sigma^{(\kappa_i + \kappa_j)}(\tau_i - \tau_j), \quad 1 \leq i, j \leq n$$

and where $\sigma^{(\kappa_i)}$ denotes the partial derivative of $\sigma(\tau_i - \tau_j)$ of order κ_i .

It can be shown that the problem of predicting $Y(t_0)$ at some new points $\tau_0 \in \mathcal{R}$, reduces to find a predictor of the form

$$\hat{Y}(\tau_0) = \sum_{k=1}^M a_k u_{k0} + \sum_{i=1}^n b_i \sigma(\tau_0, \tau_i). \quad (3)$$

where a_k and b_i are appropriate parameters to be estimated.

Note that this framework allows to use derivative data to predict derivatives [1]. Estimating derivatives and representing the dynamics for functional data is often crucial as they can reveal patterns in a (functional) dataset that address important research questions. The proposed framework also leads to the notion of *derivative principal component analysis*, which complements functional principal component analysis, one of the most popular tools of functional data analysis.

References

1. Mardia K.V., Kent, J. T., Goodall C.R.: Kriging and Splines with Derivative Information. *Biometrika* **83**, 217–285 (1996)
2. Root W. L.: Singular Gaussian measures in detection theory. In: *Proc. Symposium on Time Series Analysis*, pp. 292–315. John Wiley, New York (1963)

Two-fold data streams dimensionality reduction approach via FDA

Un approccio a due fasi per la riduzione di dimensionalità di data streams via FDA

F. Fortuna, T. Di Battista and S.A. Gattone

Abstract The analysis of big data streaming is of main interest in several fields, as it represents an important source of information. However, this type of data involves a series of challenges concerning dimensionality issues. Thus, it becomes crucial to extract relevant and reliable information from big data. To this end, we analyze data streams in a functional framework, by proposing a two-fold dimension reduction procedure. First, streams are divided into user-defined time windows, identifying for each of them a suitable probability distribution function, which is able to describe the main characteristics of the data. Second, each probability density function is reduced to a vector of functional principal component scores in order to apply standard techniques of multivariate analysis with different analysis purposes.

Abstract *L'analisi di data streams sta acquistando rilevanza in diversi settori, in quanto rappresenta un'importante fonte di informazioni. Tuttavia, questi dati comportano una serie di sfide riguardanti soprattutto problemi di dimensionalità. Diventa fondamentale, pertanto, estrarre informazioni rilevanti da una grande massa di dati. A tal fine, proponiamo di analizzare i data streams in un contesto funzionale, implementando una procedura di riduzione di dimensionalità a due step. In primo luogo, gli streams sono suddivisi in finestre temporali, identificando per ciascuna di queste un'adeguata funzione di distribuzione di probabilità, in grado di descrivere le principali caratteristiche dei dati. Nel secondo step, ogni densità di probabilità è ridotta ad un vettore di scores derivanti dall'analisi in componenti principali funzionali. Ciò permette di applicare tecniche standard di analisi multivariata con diversi scopi di analisi.*

F. Fortuna

“G.d’Annunzio” University, Pescara, Italy, e-mail: francesca.fortuna@unich.it

T. Di Battista

“G.d’Annunzio” University, Pescara, Italy, e-mail: dibattis@unich.it

S.A. Gattone

“G.d’Annunzio” University, Pescara, Italy, e-mail: gattone@unich.it

Key words: data streams, dimensionality reduction, probability density function, FPCA

1 Introduction

In recent years, the development of real time measurement instruments and data storage resources has allowed to record huge amounts of temporally ordered data, which are called data streams or big data. An important issue related to data streams is the so-called “curse of dimensionality”, which requires unbounded computational resources to uncover actionable knowledge patterns. Thus, extracting relevant and reliable information from these data becomes a crucial aspect. To address this issue, several authors [Fortuna et al., 2018, Hadjipantelis and Muller, 2018, Romano et al., 2011] have proposed to analyze data streams through the functional data analysis (FDA) approach [Ramsay and Silverman, 2005, Ferraty and Vieu, 2006]. Indeed, since functional data are infinite-dimensional objects, they provide a more suitable data streams representation than traditional multivariate vectors [Hadjipantelis and Muller, 2018]. Moreover, unlike traditional multivariate methods, FDA enables an effective statistical analysis when the number of variables exceeds the number of observations [Kosiorowski1, 2014] and the data are recorded over an irregular or even sparse design of points [Hadjipantelis and Muller, 2018].

In this paper, we propose a two-step procedure for data streams reduction based on functional data representation. In the first step, data streams are divided in user-defined and non overlapping time windows and, for each of them, a probability density function (pdf) is estimated. Indeed, the pdf allows to reduce the high dimension of the streams while retaining the most important and useful information because it express the distributional properties of large data. Although pdfs are useful tools for extracting relevant information from data streams, they are still infinite dimensional objects. Thus, in the second step of the proposed procedure, functional principal component analysis (FDA) is used to reduce the dimensionality of pdfs. Then, standard multivariate techniques can be applied to the reduced set of principal component scores. In particular, since replications of density estimates are observed over time windows, this paper focuses on the forecasting problem for density functions in a functional framework.

The remainder of the paper is organized as follows: Section 2 deals with the problem of estimating a pdf given some point evaluation of it; Section 3 revises FPCA as a dimensionality reduction method for density functions; Section 4 presents the functional time series forecasting procedure applied to pdfs of data stream time series.

2 Probability density functions in a functional framework

Probability density functions (pdfs) are frequently used to characterize the distributional properties of large-scale database systems [Hron et al., 2016] as they provide more information than using few summary statistics like totals, mean and variance [Sen and Ma, 2015]. Due to their nature, pdfs can be analyzed through the FDA approach; however, they represent a special case of functional data in that they are nonnegative everywhere and present a constant integral constraint equal to one.

Let $f_i(x)$, $i = 1, 2, \dots, n$, be a sample of n random probability density functions whose support is a finite interval $\mathcal{I} = [a, b]$, and $\mathcal{F}(\mathcal{I})$ be the function space of pdfs on \mathcal{I} defined as follows:

$$\mathcal{F}(\mathcal{I}) = \left\{ f : \mathcal{I} \rightarrow \mathbb{R}, \text{ such that } f(x) \geq 0, \text{ and } \int_{\mathcal{I}} f(x) dx = 1 \right\} \quad (1)$$

where $\mathcal{F}(\mathcal{I}) \in L^1(\mathbb{R})$, which is a convex but non linear space. This leads to difficulties for the application of common FDA methodologies, such as functional principal components analysis (FPCA), which require functions belonging to the linear space $L^2(\mathcal{I})$ [Petersen and Muller, 2016, Delicado, 2011].

A layer of complexity is added by the fact that, in real cases, data are not directly observed in the form of densities, nevertheless, for each $f_i(x)$, $i = 1, 2, \dots, n$, we observe an *i.i.d.* sample of data, y_{il} , $l = 1, \dots, L$, that are generated by the underlying distribution. Thus, the densities themselves must first be estimated ensuring that the density estimator, $\hat{f}_i(x)$, be itself a pdf, that is $\hat{f}_i(x) \geq 0$ and $\int_{\mathcal{I}} \hat{f}_i(x) dx = 1$.

The problem of obtaining a constrained estimation of the underlying function given some point-wise evaluations of it has been tackled in the literature for many decades. In this paper, we adopt the approach proposed by Ramsay [2000], according to which the constrained estimation problem is redefined into an unconstrained one by using a differential equation method. In this context, a pdf can be expressed as the solution of the first-order homogeneous differential equation:

$$Df(x) = w(x)f(x) \quad (2)$$

where $w(x) = Df(x)/f(x) = D \ln f(x)$ is itself a function. The solution of differential equation in (2) is equal to:

$$f(x) = C \exp \int_{\mathcal{I}} w(u) du = C \exp W(x) \quad (3)$$

where $W(x) = \int_{\mathcal{I}} w(u) du = D^{-1}w(x)$; $w(x)$ is an unconstrained function, which defines $f(x)$ and guarantees $f(x) > 0$, and $C = \left[\int_{\mathcal{I}} \exp W(x) dx \right]^{-1}$ is a positive normalizing constant, which satisfies the constrain $\int_{\mathcal{I}} f(x) dx = 1$. Thus, $f(x)$ is obtained by applying firstly the integration and then the exponential operator to $w(x)$ [Ramsay, 2000]. Since $w(x)$ is unconstrained, it can be expanded in terms of a set of basis functions, such as B-splines, as follows:

$$w(x) = \sum_k^K c_k B_k(x) \quad (4)$$

where B_k is the k -th spline basis function defined by a particular knot sequence and order, and c_k is the k -th coefficient defining the linear combination. The function $w(x)$ and hence $f(x)$ can be estimated by a penalized maximum likelihood method [Silverman, 1986]:

$$PENMLE = \sum_{i=1}^n \ln f(x_i) + \lambda \int \left[L \ln f(x) \right]^2 dx \quad (5)$$

In this setting, densities are estimated non-parametrically, without assuming in advance one of the many well-known parametric density functions. Indeed, the only assumption concerns smoothness.

3 Dimensionality reduction of probability density functions through FPCA

Although a probability density function represents a useful tool for extracting relevant information from data streams, it is still an infinite dimensional object, which is affected by the “curse of dimensionality”. Thus, functional principal component analysis (FPCA) [Ramsay and Silverman, 2005] is adopted to provide a representation of pdfs in the low-dimensional space spanned by a finite number of principal components. Then, standard multivariate techniques can be applied to the reduced set of principal component scores. However, since pdfs lie in the L^1 space, FPCA may lack the defining characteristics of a density leading to erroneous conclusions. For this reason, FPCA is performed on the unconstrained function $w(x)$ in Eq. (4), which belong to the Hilbert space $L^2(\mathcal{J})$ of square integrable functions with the usual inner product $\langle w, g \rangle = \int_{\mathcal{J}} w(x)g(x) dx, \forall w, g \in L^2(\mathcal{J})$ and the L^2 -norm $\|w\| = \langle w, w \rangle^{1/2} < \infty$.

Let $\{w_i(x) \in L^2(\mathcal{J}) : x \in \mathcal{J}, i = 1, 2, \dots, n\}$ be a sample of functions that can be viewed as observations of a second order stochastic process $\mathcal{W} = \{\mathcal{W}(x) \in L^2(\mathcal{J}) : x \in \mathcal{J}\}$ with mean function $\mu(x) = E[\mathcal{W}(x)]$ and covariance function $\Sigma(x, s) = Cov[\mathcal{W}(x), \mathcal{W}(s)]$. FPCA relies on the expansion of the functional data in a functional basis consisting on the eigenfunctions of the covariance operator of the process $\mathcal{W}(x)$:

$$\Gamma_{\Sigma} = \int_{\mathcal{J}} \Sigma(x, s) w(x) dx \quad (6)$$

Since Γ_{Σ} is an Hilbert-Schmidt operator, the covariance function $\Sigma(x, s)$ admits the following spectral decomposition:

$$\Sigma(x, s) = \sum_{q=1}^{\infty} \lambda_q \phi_q(x) \phi_q(s) \quad (7)$$

where λ_q are the eigenvalues (in descending order) of the covariance operator and ϕ_q the corresponding orthogonal eigenfunctions. Then, at a sample level, the Karhunen-Loève (KL) expansion [Karhunen, 1947, Loeève, 1945] of $w(x)$ can be expressed as follows:

$$w_i(x) \approx \hat{\mu}(x) + \sum_{q=1}^Q \hat{\xi}_{q,i} \hat{\phi}_q(x) + e_i(x) \quad (8)$$

where $\hat{\mu}(x)$ is the estimated mean function; $\hat{\phi}_q(x)$ is the q -th estimated functional principal component, that is the q -th eigenfunction of the empirical covariance operator; $\hat{\xi}_{q,i}$ represents the q -th estimated principal component score for the i -th observation, with $E[\xi_q] = 0$, $Var[\xi_q] = \lambda_q$, and which satisfies $\hat{\xi}_{q,i} = \int_{\mathcal{S}} (w(x) - \hat{\mu}(x)) \phi_q(x) dx$; $e_i(x)$ is a zero-mean residual function; and the expansion is truncated to a finite number of terms, $Q < n$. In this context, the tools of multivariate data analysis can be readily applied to the resulting random vector of scores, accomplishing the goal of dimension reduction.

4 Forecasting pdf via functional time series model

A streaming time series, $\{y_t\}_{t=1}^{\infty}$, is a sequence of a potentially infinite real-valued ordered observations on a discrete time grid. To solve the “curse of dimensionality” typical of big data, streaming time series are splitted into a set of windows, W_t , $t = 1, 2, \dots, T$, of a given length, such that $W_t \cap W_{t+1} = \emptyset$. The choice of the length is simple enough in practice when there is a well-defined seasonal period for which a reasonably similar behaviour is expected, such as days, weeks, months, years, etc. For each W_t , a pdf, $f_j(x)$ is estimated following the approach described in Section 2. Thus, in every W_t , we consider a curve rather than w distinctive data points. Since replications of pdf estimates over time are obtained, a functional time series (FTS) approach can be considered for forecasting purposes (see Horvath and Kokoszka [2012] for a wide review on this topic).

A FTS, $\{f_t\}_{t=1}^T$, is the realization of a functional stochastic process, i.e. a time sequence of functional observations, $f_t = \{f_t(x), x \in \mathcal{S}\}$, where t is some measurement of time and x is a continuous variable (in our case x is also a time variable) [Ramsay and Silverman, 2005, Ferraty and Vieu, 2006, Shen, 2009]. In this context, the main interest lies in forecasting future values of the time series taking into account the temporal dependence of the functional observations. In this paper, the functional forecasting problem is reduced to a multivariate version by projecting the unconstrained functions $w(x)$ in Eq. (4) into the space spanned by the Q most relevant functional principal components according to Eq. (8). Specifically, the principal component scores form the reduced dimensional multivariate dataset as they capture the dependence structure inherited in the original functional time series. Indeed, since $\hat{\xi}_{q,t}$ are uncorrelated to each other by construction, it is possible to forecast

each series $\{\hat{\xi}_{q,t}\}_{t=1}^n$ using a univariate time series model, such as the autoregressive integrated moving average (ARIMA) model of Box et al. [2008]. Alternatively, a multivariate ARIMA can be applied to the matrix of FPC scores. Conditioning on the historical curves $\mathbf{w} = \{w_1(x), \dots, w_T(x)\}$ and the fixed functional principal components $\boldsymbol{\phi} = \{\hat{\phi}_1(x), \dots, \hat{\phi}_Q(x)\}$, the h -step-ahead forecasts of $w_{T+h}(x)$ are expressed as follows:

$$\hat{w}_{T+h|T}(x) = E[w_{T+h}(x)|\mathbf{w}, \boldsymbol{\phi}] = \hat{\mu}(x) + \sum_{q=1}^Q \hat{\phi}_q(x) \hat{\xi}_{q,T+h|T} \quad (9)$$

where $\hat{\xi}_{q,T+h|T}$ denotes the h -step-ahead forecast of $\xi_{q,T+h}$ using a univariate time series model and h is a forecast horizon.

References

- G. Box, G. Jenkins, and G. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken, New Jersey, 4th ed. edition, 2008.
- P. Delicado. Dimensionality reduction when data are density functions. *Computational Statistics & Data Analysis*, 55:401–420, 2011.
- F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer, New York, 2006.
- F. Fortuna, F. Maturo, and T. Di Battista. Clustering functional data streams: Un-supervised classification of soccer top players based on Google trends. *Quality and Reliability Engineering International*, page Article in press, 2018. doi: 10.1002/qre.2333.
- P. Hadjipantelis and H. Muller. Functional data analysis for big data: A case study on california temperature trends. In W. Hardle, H. Lu, and X. Shen, editors, *Handbook of Big Data Analytics*, pages 457–483. Springer, Cham, 2018.
- L. Horvath and P. Kokoszka. *Inference for Functional Data with Applications*. Springer, New York, 2012.
- K. Hron, A. Menafoglio, M. Templ, K. Hruzova, and P. Filzmoser. Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis*, 94:330–350, 2016.
- K. Karhunen. Über lineare methoden in der "Wahrscheinlichkeitsrechnung". *Ann Acad Sci Fennicae Ser AI Math-Phys*, 37:1–79, 1947.
- D. Kosiorowski1. Functional regression in short-term prediction of economic time series. *Statistics in Transition - new series*, 15(4):611–626, 2014.
- M. Loeève. Fonctions aléatoires de second ordre. "C R" *Acad Sci Paris*, pages 220–649, 1945.
- A. Petersen and H. Muller. Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 32(1):1–1, 2016.

- J. Ramsay. Differential Equation Models for Statistical Functions. *The Canadian Journal of Statistics*, 28:225–240, 2000.
- J. Ramsay and B. Silverman. *Functional Data Analysis, 2nd edn.* Springer, New York, 2005.
- E. Romano, A. Balzanella, and L. Rivoli. Functional boxplots for summarizing and detecting changes in environmental data coming from sensors. In *Electronic Proceedings of Spatial 2, Spatial Data Methods for Environmental and Ecological Processes*. 2nd edition, Foggia 1-3 September, 2011.
- R. Sen and C. Ma. Forecasting density function: application in finance. *Journal of Mathematical Finance*, 5:433–447, 2015.
- H. Shen. On modeling and forecasting time series of smooth curves. *Technometrics*, 51:227–238, 2009.
- B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

Statistical analysis of Sylt's coastal profiles using a spatiotemporal functional model

Rik Gijsman, Philipp Otto, Torsten Schlurmann, Jan Visscher

Abstract This paper outlines a project on statistical modeling of coastal profiles. The objectives of the project are to evaluate on the morphological evolution of a coastline as well as to identify behavior of nourishments on different temporal and spatial scales. We propose the use of a flexible, spatiotemporal model for functional data, which can be estimated by an efficient algorithm based on the maximum-likelihood principle. The model is applied to measured coastal profiles of the island of Sylt in the German Bight (SE North Sea) between 1980 and 2017. Location and timing of the measurements are strongly related to the placement of nourishments. These are normally placed on a yearly basis at different locations along the island, depending on erosion losses in the previous year. Although profiles are measured on average 1-2 times per year, the spatial, temporal and functional support is heterogeneous. External forcing conditions (e.g. wave heights) that cause changes in the coastal profiles are measured in front of Sylt and could be included in the statistical model as covariates. Furthermore, change points could be assigned to profiles that are abruptly changed by a nourishment itself. The described model supports the study of coastal profile changes on other spatiotemporal scales than the measurements alone provide. An introduction into the methodology, how the model could be applied to the data set and a research outlook are provided in this study.

Key words: Coastal morphology, Functional data, Statistical modeling, Spatiotemporal modeling

1 Introduction

The 32 km long sandy westcoast of Sylt (Germany) is a morphologically active area. The island is located between two tidal inlets and is classified as an open-sand

Philipp Otto
Leibniz University Hannover, e-mail: otto@ikg.uni-hannover.de

system that continuously loses sand to the Wadden Sea (Dette, 1998). Based on coastal retreat rates between 1872 and 1984, erosion losses of 1.2 mio. cubic meters of sand are estimated (Ahrendt, 2001). The orientation of the island in combination with the wave climate result in yearly-averaged sand losses of approximately 1.0 mio. and 0.5 mio. cubic meters on the northern and the southern half of the island, respectively (Kroon et al., 2008).

In order to reduce sand losses, coastal protection measures were built along the island between 1867 and 1969 (Ahrendt and Köster, 1996). Since these measures did not reduce erosion losses, the strategy of the coastal authority of Schleswig Holstein State in Germany (*Landesbetrieb für Küstenschutz Schleswig Holstein* or *LKN.SH*) changed towards the use of nourishments in 1972. These nourishments aim at mitigating the effects of coastal erosion on the shoreline itself. This is the current state-of-the-art solution for the protection of sandy shorelines against the impacts of sea level rise (Stive et al., 2013).

Nourishments are traditionally done between the winter season, when the majority of the erosion occurs, and the summer season, when many tourists visit the beaches of Sylt. The location and design of the nourishments are determined by the *LKN.SH* after a measurement campaign along the island, and well-documented in annual reports (LKN.SH, 2016). Nourished parts of the island are measured again after the nourishment campaign. Since nourishments can be placed on the emerged and/or submerged part of the beach, these measurements normally do not cover the complete coastal transect. Next to the nourishment-related measurements, the *LKN.SH* measures the complete island with a temporal interval of approximately 3-5 years. Consequently, this has resulted in a heterogeneous spatiotemporal data set between 1972 and 2017.

The aim of this project is to develop a statistical model of the coastal profiles along the westcoast of Sylt. We initially focus on the data of a 6 km long coastal stretch on the southern half between 1980 and 2017. In particular, we consider a hierarchical model for functional spatiotemporal data, which is estimated by a spatiotemporal Expectation-Maximization (STEM) algorithm (cf. Fasso and Finazzi 2013). This algorithm has been computationally implemented by Finazzi and Fasso (2014) and Cameletti (2015). Whereas classical methods for analyzing functional data require regularly spaced observations (cf. Yao et al. 2005), the coastal profiles are characterized by a heterogenous spatial and functional support. This issue is addressed by B-spline interpolations of the functions in combination with spatial kriging techniques. Delicado et al. (2010) provide an overview on functional data analysis, if the data are correlated in space. Furthermore, they review some recent contributions. Specifically, such spatial functional data models are applied for modeling environmental data (e.g., Fasso et al. 2014; Ignaccolo et al. 2015).

The paper is structured as follows. In the next section, we sketch the statistical model and discuss several aspects of modeling spatiotemporal functional data. Thereafter, a detailed overview of the data is given in Sect. 3. Finally, Sect. 4 concludes the paper and provides an outlook for future research.

2 Statistical Model

Let $\dot{y}_s : \mathbb{Z} \times \mathbb{R}^+ \rightarrow \mathbb{R}$ be a function describing the vertical elevation z relative to a reference mean sea level, i.e., 'NN' or 'Normal Null' in Germany, which is associated to a certain distance d from a reference point s at the shoreline and a time point t . Further, $y_{s,t} = \{\dot{y}_s(d, t) : d \in D\}$ denotes the image of \dot{y}_s at a certain point of time t with the reference point s , where D is the set of all possible distances. Below, we assume that D is discrete set $\{d_1, \dots, d_k\}$. Furthermore, the profiles are observed for all stations s_1, \dots, s_n and points in time $t = 1, \dots, T$. It is worth noting that the stations are not necessarily the same at each point of time t . Furthermore, the support of the observed profiles D depends on both the reference point $s = (s_1, s_2)'$ and the point of time t .

We suppose that this functional random process follows a hierarchical functional model, i.e.,

$$y_{s,t} = \mu_{s,t} + \omega_{s,t} + \varepsilon_{s,t}, \quad (1)$$

where $\mu_{s,t}$ is the functional mean model and $\omega_{s,t}$ is a random effect model. The random errors $\varepsilon_{s,t}$ are assumed to be independent and identically distributed with zero mean and covariance matrix $\sigma_\varepsilon^2 I_k$, where I_k stands for the k -dimensional identity matrix.

For a discrete set D and $j = 1, \dots, k$, the mean model can be specified in a linear form as

$$\mu_{s,t}^{(j)} = X_{s,t}^{(j)} \beta_j \quad (2)$$

with a matrix of external regressors $X_{s,t}^{(j)}$ associated with distance d_j and a corresponding vector of coefficients β_j . Further, the random effect model is given by

$$\omega_{s,t}^{(j)} = \sum_{l=1}^m \phi_l^{(j)} z_{l,s,t}^{(j)} + \sum_{r=1}^o \alpha_r x_{r,s,t}^{(j)} w_{r,s,t}^{(j)} \quad \text{and} \quad z_{s,t}^{(j)} = G z_{s,t-1}^{(j)} + \eta_{s,t}^{(j)}, \quad (3)$$

where $z_{s,t}^{(j)} = (z_{1,s,t}^{(j)}, \dots, z_{m,s,t}^{(j)})'$ is a purely temporal component. Since the profiles are observed at irregular distances D , which may also depend on s and t , the random fields $z_{s,t}^{(j)}$ are interpolated by a B -spline approach. However, any other interpolation method could also be used. Furthermore, either time- or space-varying covariates $x_{1,s,t}^{(j)}, \dots, x_{o,s,t}^{(j)}$ could be included in the space-time component. This second term includes the spatially correlated random field $w_{r,s,t}^{(j)}$ with zero mean and unit variance. If there are no such covariates, the model collapses to

$$\omega_{s,t}^{(j)} = \sum_{l=1}^m \phi_l^{(j)} z_{l,s,t}^{(j)}, \quad (4)$$

which is commonly used in environmental studies (e.g., Finazzi and Fasso 2014; Fasso et al. 2014). Regarding this simpler model, the spatial correlation can be mod-

eled via $\eta_{s,t}^{(j)}$. That is, a certain parametric structure of the covariance matrix of $\eta_{s,t}^{(j)}$ is assumed, which then could be estimated along with all other parameters.

To model the spatial dependence in each of the j sections, isotropic or anisotropic covariance functions could be implemented. In the specific setting of this study, an anisotropic covariance function seems to be more appropriate, as the sand transport and, therefore, the profile elevation $y_{s,t}$ are affected by time-varying water levels c_{1t} , wave heights c_{2t} , periods c_{3t} and direction c_{4t} .

The model parameters can be estimated by the maximum-likelihood approach (cf. Fasso and Finazzi 2013). As the numerical maximization of the likelihood function might be computationally challenging, in particular for big data, an expectation-maximization algorithm can be used (cf. Cameletti 2015; Finazzi and Fasso 2014). To handle big geospatial data, the statistical model should be simple while being flexible enough (cf. Datta et al. 2016b,a). For example, the complexity of the spatial dependence could be reduced by covariance tapering, block tapering, fixed-rank kriging or similar methods. Vetter et al. (2014) compared several of these approaches for remotely-sensed CO₂ concentrations.

3 Description of the Data

To study decadal changes of coastal profiles affected by nourishments, we analyze 5687 coastal profiles at $n = 120$ reference points and $T = 99$ points in time along the westcoast of Sylt (Germany), see Fig. 1. This paper specifically focuses on a 6 km long coastal stretch, Puan Klent, on the southern half of Sylt. The alongshore location of the historically placed nourishments is presented as well, where the emerged (beach) and submerged (shoreface) nourishments are indicated in black and blue, respectively. The distance from the island indicates the timing of the nourishment in a window between 1972 and 2017. Pre-defined coastal transects are located with average alongshore intervals dy of 50 meter. Between 1980 and 2017, the $n = 120$ coastal transects in the study area were measured approximately 1.3 times per year. Since the measurement strategy is largely dependent on the placement of nourishments, and the nourishment locations vary in longshore direction s_1 and cross-shore direction s_2 (see Fig. 1), a heterogeneous spatiotemporal coverage is obtained.

In Fig. 2, a functional boxplot is shown for all profiles in the study area (cf., Sun and Genton 2011), where the outer thin lines represent the 99% and 1% quantiles, respectively. The central bold line shows the median and the box between the lower and upper quartiles is printed in gray. Notably, the emerged part of the coastal profiles, i.e., left of the dashed line, show larger variability. This part of the profile is represented by a milder sloping beach ($d < 40\text{dm}$) and a steeper sloping dune ($40\text{dm} < d < 200\text{dm}$). The reference point s is defined as the location of $(0, 0)'$ in \dot{y} . In addition, we plot several randomly chosen profiles to show the variability of the profiles. They do not only differ in terms of reference location, but also their length and therefore spatial coverage differs. Finally, nearshore wave conditions can cause measurements to not be on a straight line on the submerged part of the profile. The

profiles are therefore projected to a pre-defined straight coastal transect through the point s and orthogonal to the shoreline.

Furthermore, local wave conditions in front of Sylt are recorded. Three wave buoys are located in front of the south, centre and north of Sylt. In Fig. 1, their position is shown by the green circles. These buoys have determined spectral wave heights c_{2t} , periods c_{3t} and directions c_{4t} for certain periods in time. Water level measurements c_{1t} are available from a tidal gauge in front of Westerland, see the green triangle in Fig. 1. These hydrodynamic forcing conditions could also be included in the model as covariates. Since the locations and timing of the nourishments is known, and these nourishments can be interpreted as statistical change points, we can include these by not assuming any temporal dependence. Hence,

$$z_{s,t}^{(j)} = \eta_{s,t}^{(j)},$$

if the profile s was affected a nourishment at time point $t - 1$.

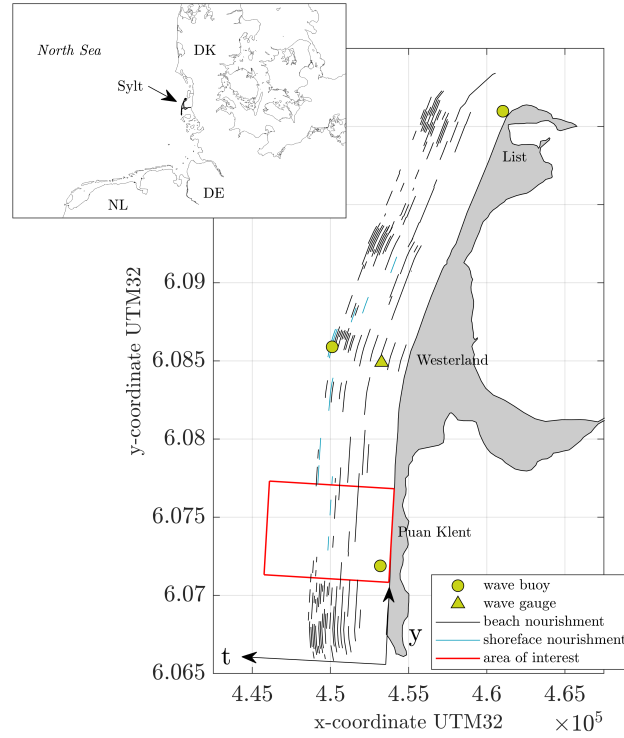


Fig. 1 Overview on beach nourishments from 1980 and 2017

4 Summary and Outlook

In this paper, we sketch how spatiotemporal functional data could be modeled using hierarchical models and the D-STEM algorithm for estimation. In particular, we consider functional profiles of the vertical elevation relative to mean sea level at the coastline of Sylt in the German Bight. These profiles were recorded between 1980 and 2017 at various locations along the island on an irregular basis. That is, the spatial domain as well as the support of the profiles are varying over time. The profiles were recorded by the *LKN.SH* to assess beach volumes and requirement of nourishments along Sylt.

Future research in the described project will focus on the following aspects. Firstly, the applicability of the model is investigated based on the profiles of the study area. Since coastal profiles can be relatively dynamic, it needs to be investigated if inclusion of the effects of the covariates improves the model capabilities. Secondly, the model is applied to the total of 37649 profiles ($T = 405$ time points and $n = 691$ locations) along the complete island of Sylt. Thirdly, long and short-term coastal profiles changes are evaluated with the statistical model. This approach could be a valuable method to increase the potential of measured shorelines worldwide.

References

- Ahrendt K (2001) Expected effect of climate change on sylt island: results from a multidisciplinary german project. *Climate Research* 18(1-2):141–146
- Ahrendt K, Köster R (1996) An artificial longshore bar at the west coast of the island of Sylt/German Bight: first experiences. *Journal of Coastal Research* pp 354–367
- Cameletti M (2015) Stem: Spatio-temporal EM. R package version 1.0
- Datta A, Banerjee S, Finley AO, Gelfand AE (2016a) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association* 111(514):800–812
- Datta A, Banerjee S, Finley AO, Hamm NAS, Schaap M (2016b) Nonseparable dynamic nearest neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Ann Appl Stat* 10(3):1286–1316, DOI 10.1214/16-AOAS931, URL <http://dx.doi.org/10.1214/16-AOAS931>
- Delicado P, Giraldo R, Comas C, Mateu J (2010) Statistics for spatial functional data: some recent contributions. *Environmetrics: The official journal of the International Environmetrics Society* 21(3-4):224–239
- Dette HH (1998) Management of beach nourishment in an open sand system. *Coastal Engineering Proceedings* 1(26)
- Fasso A, Finazzi F (2013) A varying coefficients space-time model for ground and satellite air quality data over europe. *Statistica & Applicazioni, Special Online Issue* pp 45–56

- Fasso A, Ignaccolo R, Madonna F, Demoz B, Franco-Villoria M (2014) Statistical modelling of collocation uncertainty in atmospheric thermodynamic profiles. *Atmospheric Measurement Techniques* 7(6):1803–1816
- Finazzi F, Fasso A (2014) D-STEM: a software for the analysis and mapping of environmental space-time variables. *Journal of Statistical Software* 62(6):1–29
- Ignaccolo R, Franco-Villoria M, Fasso A (2015) Modelling collocation uncertainty of 3d atmospheric profiles. *Stochastic environmental research and risk assessment* 29(2):417–429
- Kroon A, Larson M, Möller I, Yokoki H, Rozynski G, Cox J, Larroude P (2008) Statistical analysis of coastal morphological data sets over seasonal to decadal time scales. *Coastal Engineering* 55(7):581 – 600, DOI <https://doi.org/10.1016/j.coastaleng.2007.11.006>, human Interaction with Large Scale Coastal Morphological Evolution
- LKNSH (2016) Fachplan Küstenschutz Sylt (in German)
- Stive MJ, de Schipper MA, Luijendijk AP, Aarninkhof SG, van Gelder-Maas C, van Thiel de Vries JS, de Vries S, Henriquez M, Marx S, Ranasinghe R (2013) A new alternative to saving our beaches from sea-level rise: The sand engine. *Journal of Coastal Research* pp 1001–1008, DOI 10.2112/JCOASTRES-D-13-00070.1
- Sun Y, Genton MG (2011) Functional boxplots. *Journal of Computational and Graphical Statistics* 20(2):316–334
- Vetter P, Schmid W, Schwarze R (2014) Efficient approximation of the spatial covariance function for large datasets - analysis of atmospheric co2 concentrations. *Journal of Environmental Statistics* 6(3):1–36, URL <http://jes.stat.ucla.edu/v06/i03>
- Yao F, Müller HG, Wang JL (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100(470):577–590

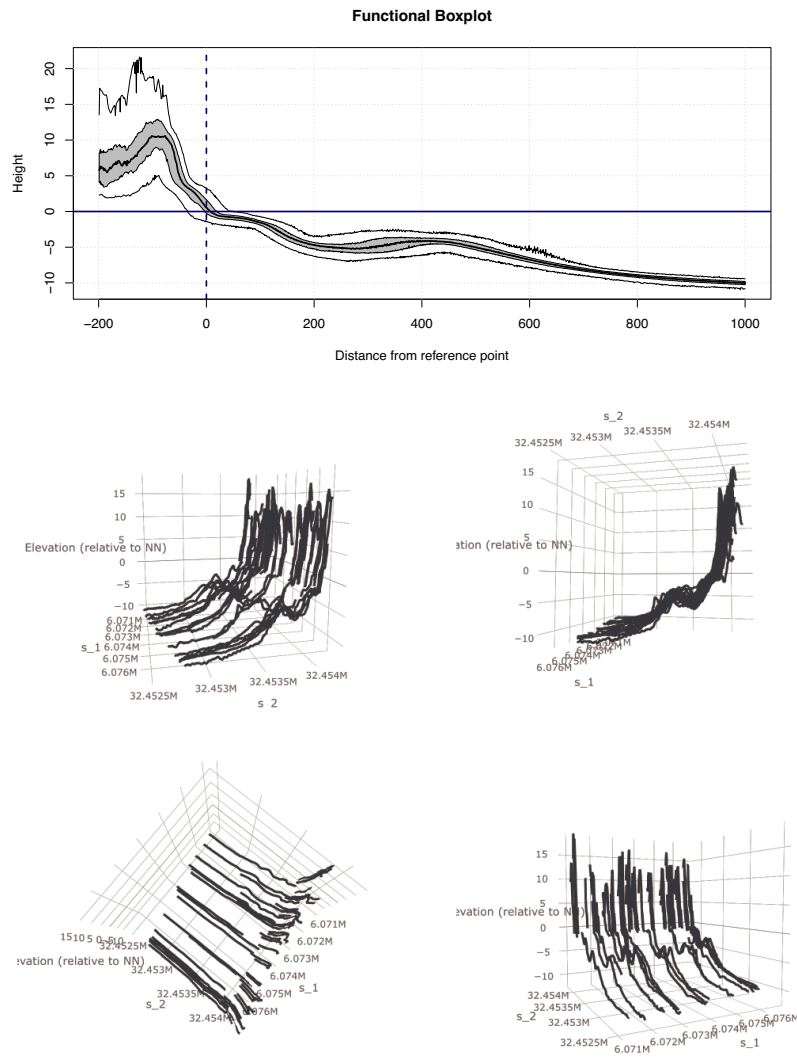


Fig. 2 Functional boxplot (above) and exemplary profiles (below)

Bootstrap prediction intervals for weighted TAR predictors

Intervalli di previsione bootstrap per previsori ponderati per modelli TAR

Francesco Giordano and Marcella Niglio

Abstract We evaluate the forecast accuracy of a new predictor proposed for the Self Exciting Threshold AutoRegressive (SETAR) model. In particular we focus the attention on a new predictor obtained as weighted mean of the past observations, whose weights are obtained from the minimization of the Mean Square Forecast Errors. Even if the “point accuracy” of this weighted predictor has been performed, the study of its distribution and in particular the construction of the prediction intervals has not been faced. Starting from the evaluation that the prediction errors, obtained from the difference between the true future values and the predicted values, follow a nonstandard distribution, in this contribution we focus the attention on different bootstrap methods for dependent data that allow to construct prediction intervals for the weighted SETAR predictor and their coverage is properly compared.

Abstract *Il contributo si pone l'obiettivo di valutare l'accuratezza dei previsori generati da modelli per serie storiche denominati "a soglia". In tale contesto l'attenzione sarà posta su un nuovo previsore proposto per tali modelli a soglia, ottenuto dalla media ponderata delle osservazioni. Più in dettaglio i pesi sono generati dalla minimizzazione della media del quadrato degli errori di previsione. L'accuratezza di tali previsori puntuali è stata già esaminata, ma la loro distribuzione e la costruzione di intervalli di previsione, non è stata ancora affrontata. Per tale motivo l'attenzione sarà rivolta alla costruzione di intervalli di previsione, fondata sull'uso di approcci bootstrap, di cui è opportunamente valutata la copertura.*

Key words: Threshold model, prediction intervals, coverage

Francesco Giordano

Department of Economics and Statistics, Università degli Studi di Salerno, Italy, e-mail: giordano@unisa.it

Marcella Niglio

Department of Economics and Statistics, Università degli Studi di Salerno, Italy e-mail: mniglio@unisa.it

1 Introduction

One of the main aim of the time series analysis is the forecasts generation. Despite the high interest that this topic has always raised in the literature, the attention is often focused on the generation of the so called “point forecasts” whereas the distribution of the predictor and the building of the prediction intervals (PI in the following) do not receive similar attention. The reason can be found in the difficulty to obtain “analytic” results on the distribution of the predictors that, as will be discussed in the following, can be given only under well defined and simplified conditions. Even though these difficulties, the advantages that can be obtained from the construction of the prediction intervals are unchallengeable: they allow to define the interval in which a future value belongs with a given probability; they allow to evaluate the variability (and then the uncertainty) related to the selected predictor; they allow to consider the full distribution of the predictor and not only the single value obtained from the generation of point forecasts; they give, even to the practitioners, a clear information on how confident they should be with the generated prediction.

As expected the building of the PI (as the generation of the point forecasts) is characterized by a different level of complexity that grows as the complexity of the time series model increases. Note that when a new proposal is given in this domain, it is needed to face a clear trade-off: to guarantee the accuracy of the method (properly evaluated, for example in terms of PI coverage); to prevent the introduction of strict assumptions on the generating process that definitely limit the use of the proposal. In this context, starting from the contributions of [3] and [4] that summarize the main approaches to build PI's and given the results on the linear and nonlinear time series domain differently documented in [10], [9] and, in particular, in [8] (to cite the more recent reference books), in the present paper, in Sect. 2, we discuss the foundations of the PI's building in time series analysis starting from the case where strong conditions are given on the generating process that are then gradually removed. In Sect. 3 the attention is, instead, focused on the construction of PI's for a new predictor proposed in nonlinear time series domain. Given the complexity of the generating process, the PI si obtained using a bootstrap approach whose performance has been evaluated through a Monte Carlo study. Final remarks are given at the end.

2 Prediction intervals: from the linear to the nonlinear domain

Before to present how the PI's can be obtained in the nonlinear time series domain, it is useful to understand how, and under which conditions, they can be build in the linear context even highlighting how the removal of some assumptions change definitely the theoretical foundations of the problem.

The construction of the prediction intervals in the linear time series domain can be faced considering a growing level of generality (which is obviously connected to a decreasing number of assumptions on the generating process).

Let X_t be a stationary autoregressive model of order p :

$$X_t = \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} + a_t \quad (1)$$

with $a_t \sim \text{i.i.d.}$ random variables, with $E[a_t] = 0$ and $E[a_t^2] = \sigma_a^2$, it is widely known that under proper assumptions on the parameters [2], the process (1) can be written as an infinite weighted sum of innovations a_t 's:

$$X_t = 1 + \sum_{j=1}^{\infty} \psi_j a_{t-j} \quad (2)$$

whereas the best linear h -steps ahead predictor (that minimizes the mean square forecast error) is given by:

$$\hat{X}_{t+h} = \phi_1 \hat{X}_{t+h-1} + \dots + \phi_p \hat{X}_{t+h-p} \quad (3)$$

with $\hat{X}_{t+h-p} = X_{t+h-p}$ if $h \leq p$.

Under the assumption that: the autoregressive order p and the parameters ϕ_i , for $i = 1, \dots, p$, of the model are all known, the innovations $\{a_t\}$ are a sequence of independent Gaussian random variables, the $(1 - \alpha)\%$ PI for X_{t+h} is given by:

$$\hat{X}_{t+h} \pm z_{(1-\alpha/2)} \sigma_a^2 \left(1 + \sum_{j=1}^{h-1} \psi_j^2 \right)^2 \quad (4)$$

with $z_{(1-\alpha/2)}$ the $(1 - \alpha/2)$ quantile of the standard Gaussian random variable.

Note that the result (4) is based on well defined assumptions:

- (A.1) the generating process X_t is linear (and autoregressive);
- (A.2) the distribution of the innovations is Gaussian;
- (A.3) the parameters are known (and fulfill some conditions given in [2]).

The assumptions (A.1)-(A.3), combined with the further assumptions given before on the distribution of the innovations, guarantee the Gaussian distribution of the predictor but, at the same time, does not consider the variability induced by the estimation of the model parameters.

The removal of one or more of these assumptions leads to a non standard distribution of \hat{X}_{t+h} and so the construction of the PI's needs to be completely revised.

In the linear domain a large number of proposals has been given in the literature where the assumptions (A.2) and (A.3) are weakened and the main consequences on the predictor distribution are differently faced.

Among them the methods based on the bootstrap approach have received large attention and they are mainly focused on three proposals based on the issues given in [13]. In the autoregressive domain the main proposals are given in: [24], [18] and [12], [7], [21], [14], [15], [16] whereas, more recently and in a wider autoregressive domain, in [20]. The cited proposals have differently faced the construction of the bootstrap PI's for linear autoregressive models evaluating: the impact of the parameters estimation, the removal of the Gaussianity of the innovations, the effects

of the estimation methods, the bias of the parameters estimator in presence of small samples, the coverage of the PI's when different approaches are considered to generate the bootstrap series (for example based on backward or forward procedures, with centered or standardized bootstrap innovations etc.).

Starting from these encouraging results developed in the linear autoregressive domain, different proposals to build bootstrap PI's are given when even the assumption (A.1) is removed.

In more detail, in the following our attention will be focused on the PI's of a nonlinear class of autoregressive structures, called Self Exciting Threshold AutoRegressive (SETAR) models ([25], [26]) that, despite their structure is locally linear, have a dynamic structure that cannot take advantage of the results developed in the linear domain (even when the assumption (A.2) and (A.3) are considered).

Let X_t , with $t \in T$, be a nonlinear SETAR(2; p_1, p_2) model, its stochastic structure is given by:

$$X_t = \sum_{j=1}^k \left(\phi_0^{(j)} + \phi_1^{(j)} X_{t-1} + \dots + \phi_{p_j}^{(j)} X_{t-p_j} \right) \mathbb{I}_{\{X_{t-d} \in \mathbb{R}_j\}} + \varepsilon_t, \quad (5)$$

where k is the number of regimes, p_j is the autoregressive order of regime j , $\mathbb{I}_{\{\cdot\}}$ is an indicator function, X_{t-d} is the *threshold variable*, d is the *threshold delay*, \mathbb{R}_j is a subset of the real line such that $\mathbb{R}_j \cap \mathbb{R}_{j'} = \emptyset$, for $j \neq j'$ and $\bigcup_{j=1}^k \mathbb{R}_j = \mathbb{R}$, with $\mathbb{R}_j = (r_{j-1}, r_j]$ and $-\infty = r_0 < r_1 < \dots < r_{k-1} < r_k = \infty$, with r_j the *threshold value*, for $j = 1, \dots, k$.

The generation of point forecasts for model (5) has been differently faced (see among the others [5], [6], [1]) and the main results on the generation of the prediction intervals based on bootstrap approaches has been recently given in [17], [23] and [20]. Note that in all cited contributions the predictor considered for model (5) is given by conditional expectation of X_{T+h} . In other words $\hat{X}_{T+h} = E[X_{T+h} | \mathbf{X}_T]$, with \mathbf{X}_T the information available up to time T , such that:

$$\begin{aligned} \hat{X}_{T+h} = & \left(\hat{\phi}_0^{(1)} + \hat{\phi}_1^{(1)} \hat{X}_{T+h-1} + \dots + \hat{\phi}_p^{(1)} \hat{X}_{T+h-p+1} \right) \mathbb{I}_{\{\hat{X}_{T+h-d} \in \hat{\mathbb{R}}_1\}} \\ & + \left(\hat{\phi}_0^{(2)} + \hat{\phi}_1^{(2)} \hat{X}_{T+h} + \dots + \hat{\phi}_p^{(2)} \hat{X}_{T+h-p+1} \right) (1 - \mathbb{I}_{\{\hat{X}_{T+h-d} \in \hat{\mathbb{R}}_1\}}), \end{aligned} \quad (6)$$

with $\hat{X}_{T+h-i} = X_{T+h-i}$, if $h \leq i$, and $\hat{\mathbb{R}}_1$ is the real interval $(-\infty, \hat{r}_1]$.

In the following we consider a new predictor for model (5) and we mainly focus the attention on its variability evaluated through bootstrap prediction intervals.

3 Prediction intervals for a TAR predictor: a bootstrap approach

Even if the forecasts generation from model (5) has been differently evaluated, the predictor is always based on the conditional expectation (6). In this context a new proposal has been given in Niglio [19] that introduces a predictor for one-step ahead

forecasts from SETAR(2; p, p) models, based on the weighted average of the past observations:

$$\hat{X}_{T+1}^w = \sum_{t=\max\{p,d\}+1}^T w_t X_t \quad (7)$$

where the weights w_t , $t = \max\{p, d\} + 1, \dots, T$, are obtained from the minimization of the Mean Square Forecast Error:

$$\min_{\mathbf{w} \in \mathbf{W}} E[(X_{T+1} - \hat{X}_{T+1}^w)^2],$$

with \mathbf{W} a compact set for \mathbf{w} .

Instead of the evaluation of the point forecast (7) has been evaluated, the building of the PI's has not been faced, even if the advantage to generate these intervals has been discussed in the previous pages.

Given the nonlinearity of the generating process that makes unavailable the derivation of the "exact" distribution of \hat{X}_{T+1}^w , we take advantage of the bootstrap proposal of [20] to define the PI that for the predictor (7) can be so sketched:

PI Algorithm: Fitted Residuals

[Step 1.] Estimate the model parameters, $\hat{\phi}_i^{(j)}$ and \hat{r}_1 , $i = 1, \dots, p$, $j = 1, 2$ of model (5);

[Step 2.] Compute the fitted residuals:

for (t in $p + 1 : T$)

$$\hat{\varepsilon}_t = X_t - \hat{X}_t;$$

[Step 3.] Center the residuals:

$$\hat{\varepsilon}_t^c = \hat{\varepsilon}_t - \text{mean}(\hat{\varepsilon})$$

with $\hat{\varepsilon}$ the vector of the estimated residuals in Step 2 .

[Step 4.] Generate the bootstrap pseudo residuals:

Sample with replacement from the series of the centered estimated residuals $T - p + 1$ innovations:

for (t in $p + 1 : T$)

$$\varepsilon_t^* = \hat{\varepsilon}_t^c$$

[Step 5.] Generate the bootstrap series

select randomly a vector of p consecutive values from the times series, $(X_k, X_{k+1}, \dots, X_{k+p-1})$, for $k = 1, \dots, T - p + 1$:

$$(X_1^*, X_2^*, \dots, X_p^*) = (X_k, X_{k+1}, \dots, X_{k+p-1})$$

for (t in $p + 1 : T + m$)

$$\begin{aligned} X_t^* = & \left(\hat{\phi}_0^{(1)} + \hat{\phi}_1^{(1)} X_{t-1}^* + \dots + \hat{\phi}_p^{(1)} X_{t-p}^* \right) \mathbb{I}_{\{X_{t-1}^* \in \hat{\mathbb{R}}_1\}} \\ & + \left(\hat{\phi}_0^{(2)} + \hat{\phi}_1^{(2)} X_{t-1}^* + \dots + \hat{\phi}_p^{(2)} X_{t-p}^* \right) (1 - \mathbb{I}_{\{X_{t-d}^* \in \hat{\mathbb{R}}_1\}}) + \varepsilon_t^* \end{aligned} \quad (8)$$

and discard from the pseudo-series the first m artificial data

for (t in $1 : T$)

$$X_t^* = X_{t+m}^*$$

PI Algorithm: Fitted Residuals, (cont.)

[Step 6.] estimate the parameters of the SETAR(2; p, p) model using the pseudo-data and then compute the bootstrap weighted forecast (7) :

$$\hat{X}_{T+1}^{w*} = \hat{\mathbf{w}}^{*\top} \mathbf{X}^*$$

where the weights $\hat{\mathbf{w}}^*$, $t = p + 1, \dots, T$ are estimated from the bootstrap series;

[Step 7.] generate the future bootstrap observations

$$\begin{aligned} X_{T+1}^* = & \left(\hat{\phi}_0^{(1)} + \hat{\phi}_1^{(1)} X_T^* + \dots + \hat{\phi}_p^{(1)} X_{T-p+1}^* \right) \mathbb{I}_{\{X_T^* \in \mathbb{R}_1\}} \\ & + \left(\hat{\phi}_0^{(2)} + \hat{\phi}_1^{(2)} X_T^* + \dots + \hat{\phi}_p^{(2)} X_{T-p+1}^* \right) (1 - \mathbb{I}_{\{X_T^* \in \mathbb{R}_1\}}) + \varepsilon_{T+1}^* \end{aligned} \quad (9)$$

with $X_t^* = X_t$, for $t = T - p + 1, \dots, T$;

[Step 8.] calculate the bootstrap prediction error

$$\hat{e}_{T+1}^* = X_{T+1}^* - \hat{X}_{T+1}^{w*};$$

[Step 9.] repeat B times [Step 3.]-[Step 8.] and compute the $\alpha/2$ and $1 - \alpha/2$ quantiles, $q_{(\alpha/2)}^*$ and $q_{(1-\alpha/2)}^*$, of the empirical distribution of the bootstrap prediction error \hat{e}_{T+1}^* ;

[Step 10.] compute the one-step ahead prediction from (7), \hat{X}_{T+1}^w , using the parameters estimated in [Step 1.];

[Step 11.] construct for X_{T+1} the prediction interval:

$$[\hat{X}_{T+1}^w + q_{(\alpha/2)}^*, \hat{X}_{T+1}^w + q_{(1-\alpha/2)}^*]$$

Following [20], the *fitted residuals* algorithm can be properly modified to obtain the *predicted residuals* algorithm where the only difference is related to [Step 1] where one observation at time, X_t , for $t = p + 1, \dots, T$ is deleted from the series to estimate the parameters $\hat{\phi}_i^{(k,t)}$, $\hat{r}^{(t)}$, for $k = 1, 2$ and $i = 1, \dots, p$. Then the fitted values in [Step 2.] are obtained using the parameters $\hat{\phi}_i^{(k,t)}$, $\hat{r}^{(t)}$ and then the *predictive residuals* (that replace the fitted residuals in [Step 2.]) becomes $\hat{\varepsilon}_t^{(t)} = X_t - \hat{X}_t^{(t)}$.

4 Simulation study

To evaluate the performance of the bootstrap approach, we have made a Monte Carlo simulation study where we have considered three SETAR models:

- (M1) $X_t = 0.5X_{t-1} \mathbb{I}_{\{X_{t-1} \in \mathbb{R}_1\}} - 0.9X_{t-1} (1 - \mathbb{I}_{\{X_{t-1} \in \mathbb{R}_1\}}) + \varepsilon_t$, with $\varepsilon_t \sim N(0, 1)$
- (M2) $X_t = 0.5X_{t-1} \mathbb{I}_{\{X_{t-1} \in \mathbb{R}_1\}} - 0.9X_{t-1} (1 - \mathbb{I}_{\{X_{t-1} \in \mathbb{R}_1\}}) + \varepsilon_t$, with $\varepsilon_t \sim \text{Laplace}(0, 1)$
- (M3) $X_t = (0.5X_{t-1} - 0.2X_{t-2}) \mathbb{I}_{\{X_{t-1} \in \mathbb{R}_1\}} + (X_{t-1} - 0.8X_{t-2}) (1 - \mathbb{I}_{\{X_{t-1} \in \mathbb{R}_1\}}) + \varepsilon_t$ with $\varepsilon_t \sim N(0, 1)$.

where model M3 has been selected by considering the larger stationarity conditions, with respect to those given in the literature, for the SETAR(2; 2,2) model as

Table 1 Simulation results from $n = 500$ time series generated from the models, M1, M2 and M3, considering two levels of nominal coverage: 0.90, 0.95. CVR: empirical coverage of the PI; LEN: average length of the 500 PI's; s.e.: standard error of the 500 PI's length.

	Nominal coverage $1 - \alpha = 0.90$			Nominal coverage $1 - \alpha = 0.95$		
	CVR	LEN	s.d.	CVR	LEN	s.d.
Model M1						
Fitted Residuals	0.9006	3.9124	0.2717	0.9496	4.7353	0.3601
Predictive Residuals	0.9079	4.0664	0.3065	0.9543	4.9205	0.4067
Li [17]	0.8700	3.8460	0.2630	0.9243	4.6477	0.3434
StaWink [23]	0.8707	3.9755	0.1627	0.9238	4.7918	0.2375
Model M2						
Fitted Residuals	0.8932	5.5131	0.5802	0.9431	7.0198	0.8203
Predictive Residuals	0.8991	5.7105	0.5959	0.9480	7.2840	0.8371
Li [17]	0.8633	5.4524	0.5606	0.9200	6.9155	0.7988
StaWink [23]	0.8695	5.6064	0.3048	0.9278	7.1610	0.4284
Model M3						
Fitted Residuals	0.8933	4.1017	0.2969	0.9442	4.9387	0.3915
Predictive Residuals	0.8961	4.3050	0.3443	0.9456	5.1839	0.3424
Li [17]	0.8835	4.0429	0.2787	0.9345	4.8658	0.3424
StaWink [23]	0.8864	4.1608	0.2167	0.9344	5.0217	0.4846

proposed in [11].

For each model we have generated 500 artificial time series of length $T = 150$. For each of them we have constructed a bootstrap prediction interval (following the fitted and predictive residuals algorithms) and considering $B = 1000$ bootstrap sample paths. The bootstrap prediction intervals of the 500 time series are then evaluated considering their empirical coverage (CVR), the average length (LEN) and the standard deviation of their lengths (s.d), as described in [20, p. 13]. Finally, using the same criteria, the bootstrap PI's are compared to those proposed in [17] and [23].

It is interesting to note, from the results in Table 1, that the fitted and predictive residuals algorithms based on [20] always outperform the other two approaches in terms of empirical coverage even with the more complex structure given by model M3. Further, given the selected series length, the fitted and the predictive residuals appear to show similar performance in terms of coverage, even if the predictive approach is characterized, as expected, by a higher variability of the PI's length. The removal of the assumption of the Gaussian residuals and the introduction of a Laplace distribution for ε_t , from model M1 to model M2, do not markedly change the coverage results but increases both the average length and the variability of the PI's.

These encouraging results empirically show the consistency of the PI's proposed for the predictor (7) that will be further investigated and theoretically examined.

References

1. Boero G., Marrocu M.: The performance of SETAR models: a regime conditional evaluation of point, interval and density forecasts. *Int J Forecasting* **20**, 305-320 (2004)
2. Box, G. E. P., Jenkins, G. M., Reinsel G.C.: *Time-Series Analysis, Forecasting and Control*, (3rd ed.) San Francisco: Holden-Day (1994)
3. Chatfield C.: Calculating interval forecasts (with discussion). *J Bus Econ Stat* **11**, 121-144 (1993)
4. Chatfield C.: Prediction Intervals for Time-Series Forecasting. In: Armstrong, J.S. (Ed.) *Principles of Forecasting. A Handbook for Researchers and Practitioners*. Springer, Boston (2001)
5. Clements, M.P., Franses, J.S., Van Dijk, D.: On SETAR non-linearity and forecasting. *J Forecasting*, **22** 359-375 (2003)
6. Clements M.P., Smith J.: The performance of alternative forecasting methods for SETAR models. *Int J Forecasting* **13**, 463-475 (1997)
7. Clements, M P., Taylor N.: Bootstrapping prediction intervals for autoregressive models. *Int J Forecasting* **17**, 247-267 (2001)
8. De Gooijer, J.: *Elements of Nonlinear Time Series Analysis and Forecasting*. Springer, Switzerland (2017)
9. Elliott, G., Timmermann, A.: *Economic Forecasting*, Princeton University Press, Princeton (2016)
10. Franses, P.H., van Dijk D., Opschoor A.: *Time series models for business and economic forecasting* (Second revised edition). Cambridge University Press, Cambridge (2014)
11. Giordano F., Niglio M., Vitale C.D.: Ergodicity of the threshold autoregressive process with two regimes. Manuscript (2019)
12. Grigoletto, M.: Bootstrap prediction intervals for autoregressions: some alternatives. *Int J Forecasting* **14**, 447-456 (1998)
13. Hall P.: *The Bootstrap and Edgeworth Expansion*. Springer, New York (1992)
14. Kim J.H.: Bootstrap-after-bootstrap prediction intervals for autoregressive models. *J Bus Econ Stat* **19**, 117-128 (2001)
15. Kim J.H.: Bootstrap prediction intervals for autoregressive models of unknown of infinite lag order. *J Forecasting* **21**, 265-280 (2002)
16. Kim J.H.: Bootstrap prediction intervals for autoregression using asymptotically mean-unbiased estimators *Int J Forecasting*, **20**, 85-97 (2004)
17. Li, J. 2011: Bootstrap prediction intervals for SETAR models. *Int J Forecasting* **27**, 320-332.
18. Masarotto, G.: Bootstrap prediction intervals for autoregressions, *Int J Forecasting* **6**, 229-239 (1990)
19. Niglio, M. SETAR forecasts with weighted observations. Submitted manuscript (2019)
20. Pan, L., Politis, D. N.: Bootstrap prediction intervals for liner, nonlinear and nonparametric autoregressions. *J Stat Plan Infer* **177**, 1-27 (2016)
21. Pascual L., Romo, J., Ruiz, E.: Effects of parameter estimation on prediction densities: a bootstrap approach. *Int J Forecasting* **17**, 83-103 (2001)
22. Politis, D.: Model-free model-fitting and predictive distributions (with discussions). *Test* **22**, 183-250 (2013)
23. Staszewska-Bystrova A., Winker, P.: Improved bootstrap prediction intervals for SETAR models. *Stat Papers* **57**, 89-98 (2016)
24. Thombs L.A., Schucany W.R.: Bootstrap prediction intervals for autoregressions, *J Am Stat Ass* **85**, 486-492 (1990)
25. Tong H.: *Threshold Models in Non-linear Time Series Analysis*. Lecture Note in Statistics. Springer Verlag, New York (1983)
26. Tong H.: *Non-linear Time Series: A Dynamical System Approach*, Oxford University Press, New York (1990)

A rank graduation index to prioritise cyber risks

Un indice di graduazione per assegnare livelli di priorità ai rischi informatici

Paolo Giudici and Emanuela Raffinetti

Abstract In this paper we introduce a new methodology for estimating the risks of cyber attacks. In order to deal with the ordinal nature of the cyber risk response variable, an extension of linear regression models is proposed, by means of the rank tools. We also suggest a specific model evaluation measure, called *RG* (Rank Graduation), aiming at detecting the factors which mainly affect cyber risks. Finally, to shed light on the effectiveness of our proposal, we use our proposed methodology to rank real cyber loss data.

Abstract *In questo articolo, introduciamo una nuova metodologia per la stima dei rischi legati agli attacchi informatici. Allo scopo di superare le problematiche associate alla natura ordinale della variabile risposta, identificabile con il rischio informatico, proponiamo un'estensione dei modelli di regressione lineare basata sull'utilizzo dei ranghi. Infine, con l'obiettivo di individuare i fattori che principalmente incidono sul rischio informatico, una nuova misura di valutazione del modello, chiamata RG (Rank Graduation), viene presa in considerazione. L'articolo si conclude con un'interessante applicazione della metodologia proposta ai dati reali, che mette ulteriormente in evidenza la sua efficacia nel processo di classificazione dei rischi informatici.*

Key words: cyber risk, ordinal variables, rank-based methods

Paolo Giudici

Department of Economics and Management, University of Pavia, Via San Felice 5, 27100 Pavia (Italy), e-mail: paolo.giudici@unipv.it

Emanuela Raffinetti

Department of Economics, Management and Quantitative Methods, University of Milan, Via Conservatorio 7, 20122 Milan (Italy), e-mail: emanuela.raffinetti@unimi.it

1 Introduction

In the last few years the number of cyber attacks has surged, with a growth of about 30% between 2014 and 2017. The trend in 2018 follows a similar behavior, with 730 cyber attacks observed only in the first half of the year [2]. Cyber risks can be defined as “any risk emerging from the use of information and communication technology (ICT) that compromises the confidentiality, availability, or the integrity of data or services” (see e.g. [4]).

Financial institutions are encouraged by regulators to use statistical approaches to estimate the capital charge covering operational risk, which include cyber risks. This requires the presence of historical loss data, in a quantitative format. We remark that cyber events are typically expressed on ordinal scales. While the literature on the quantitative measurement of operational risks (see e.g. [3]), based on loss data, constitute a reasonably large body, that on cyber risk measurement and, especially, on ordinal cyber risk measurement, is very limited. Our contribution tries to fill this gap in the literature, providing a cyber risk model based on ordinal data. Specifically, given the ordinal nature of the target variable measuring the severity degree, a new approach that extends linear regression models is introduced. Furthermore, since an essential part of the cybersecurity management is to detect the main factors affecting the severity degree, it seems appropriate to validate the different models used for detecting the variables impacting on it through specific predictive accuracy measures.

Typically, the choice of the most suitable validation metric is strictly related to the nature of the response variable to be predicted. Recently, a measure that is objective and not endogenous to the system itself was suggested by [5] to evaluate the model predictive accuracy in presence of both binary and continuous response variables. In this paper, an extension of this measure to the case of discrete variables is proposed with the aim of providing a new model selection criterion when comparing different models.

The paper is organized as follows. Section 2 introduces our proposal. Section 3 illustrates the application of the proposed methodology to real data concerning cyber attacks collected at the worldwide level. Finally, the last section concludes.

2 Methodology

The proposal presented in this contribution is twofold. On the one hand, a novel model specification in the case of ordinal response variable, as is the severity variable considered in cyber risk measurement, is introduced. On the other hand, a new criterion for the comparison of different cyber risk models is illustrated.

2.1 The rank regression model

As the cyber events are typically rare and not repeatable, it is quite natural to measure them with a less demanding ordinal approach rather than using quantitative data which are often not available. Ordinal data for cyber risk measurement can be summarised, by means of a pair of statistics for each event type: the frequency of the event: how many times it has occurred, in a given period; and the corresponding severity: the mean observed loss. In the context of ordinal data, the severity can be expressed on an ordinal scale, characterised by $K = k$ distinct levels, arranged according to the corresponding magnitude. To understand the causes of cyber risks, each observed severity can be associated to a vector of explanatory variables, such as the type of attack, the technique of the attack, the victim type and the geographical area where the event has occurred.

The statistical models typically used to explain an ordinal response variable with a set of p explanatory variables are the ordered logit or probit models (see, for instance [7] and [1]). These, however, may be difficult to summarise and interpret, especially in applied contexts. We therefore develop linear regression models for a response variable that takes ordinal values. With the aim of avoiding an arbitrary assignment of the measurement scale, we resort to the ranks.

Let Y be a response variable, expressed through k ordered categories. A rank $r_1 = 1$ to the smallest ordered category of Y and a rank $(r_{j-1} + n_{j-1})$ to the following ordered categories, where n_{j-1} is the absolute frequency associated with the $(j-1)$ -th category and $j = 2, \dots, k$, are assigned. Based on this transformation, the phenomenon described by the Y variable can be re-formulated in terms of its ranks R , where:

$$R = \left\{ \underbrace{r_1, \dots, r_1}_{n_1}, \underbrace{r_2, \dots, r_2}_{n_2}, \dots, \underbrace{r_k, \dots, r_k}_{n_k} \right\}, \quad (1)$$

with $r_1 = 1$, $r_2 = r_1 + n_1$ and $r_k = r_{k-1} + n_{k-1}$.

Given p explanatory variables, a regression model for R can be specified as follows

$$\hat{r} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p, \quad (2)$$

whose unknown parameters can be estimated by the classical Maximum Likelihood method.

2.2 The RG as a criterion for model comparison

Typically, the specification of a model is completed by a procedure that compares different models and choose the best one in terms of goodness of fit. Here, we sug-

gest a novel metric, that takes into account the ordinal nature of the response variable. A similar measure, named RG , was originally provided by [5] as a criterion for the model validation in the case of both binary and continuous variables. Following [5], we extend the RG measure to the context of response ordinal variable.

Let R , defined in (1), be the vector of the rank-transformed response variable values and \hat{R} be the vector of the corresponding predicted values. The R values can be used to build the L_R Lorenz curve (see, [6]), characterised by the following pairs: $(i/n, \sum_{j=1}^i r_{ord(r_j)} / \sum_{i=1}^n r_{ord(r_i)})$, for $i = 1, \dots, n$, where $r_{ord(r_i)}$ indicates the rank-transformed response variable values ordered in a non-decreasing sense. Analogously, the R values can also be re-ordered in a non-increasing sense, providing the L'_R dual Lorenz curve.

Let $r_{ord(\hat{r}_i)}$, for $i = 1, \dots, n$, indicate the R values re-ordered according to corresponding predicted values given by the model in (2). The set of pairs $(i/n, \sum_{j=1}^i r_{ord(\hat{r}_j)} / \sum_{i=1}^n r_{ord(r_i)})$ provides the so-called C concordance curve which measures the concordance between the response variable R and the corresponding predicted variable \hat{R} orderings. In addition, the set of pairs $(i/n, i/n)$ detects the bisector curve, for $i = 1, \dots, n$, which corresponds to the case of a random model occurring if the predicted variable values are all equal each other. For the sake of clarity, a graphical representation of the four curves is given in Figure 1.

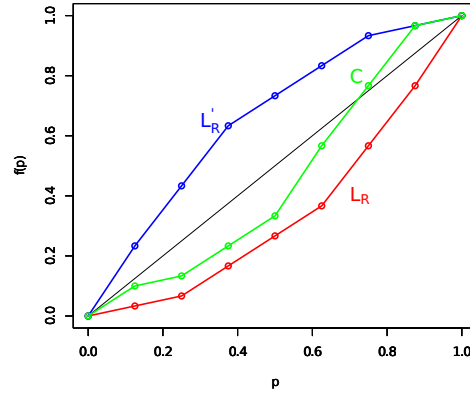


Fig. 1 The L_R (red) Lorenz curve, dual L'_R (blue) Lorenz curve, C (green) concordance curve and bisector curve (black).

The previous quantities can give rise to a new model selection tool, which is named RG as in [5]. Its formula is:

$$RG = \sum_{i=1}^n \frac{\left\{ (1/(n\bar{r})) \sum_{j=1}^i r_{ord(\hat{r}_j)} - i/n \right\}^2}{i/n}, \quad (3)$$

where \bar{r} is the mean of all ranks.

A more concise expression for RG can also be derived as follows:

$$RG = \sum_{i=1}^n \frac{\{C(r_{ord}(\hat{r}_j)) - i/n\}^2}{i/n}, \quad (4)$$

where $C(r_{ord}(\hat{r}_j)) = \frac{\sum_{j=1}^i r_{ord}(\hat{r}_j)}{\sum_{i=1}^n r_{ord}(\hat{r}_i)}$ is the cumulative values of the (normalised) rank-transformed response variable.

We remark that the RG in (3) and (4) are expressed in absolute terms. When comparing different models, a relative measure appears more appropriate making the interpretation straightforward. The relative RG version, denoted with RG_{norm} , can be specified as the ratio between its value and its maximum value, which is reached when the ordering of the rank-transformed response variable perfectly overlaps with the ordering of the corresponding predicted values. On the contrary, the RG minimum value is reached if the predicted values provided by the model are the same.

3 Application to cyber risk data

In this section we discuss an application of our proposals to cyber risk data collected by the Clusit Association, the most relevant and respected Italian association in the field of information security. The association includes, as member organisations, companies from different fields such as: Banks, Insurances, Public Administrations, Health companies and Telecommunication companies.

The data we consider consists of 6,865 worldwide observations on serious cyber attacks, in the years 2011-2017. An attack is classified as “serious” if it has led to a significant impact, in terms of economic losses and/or damages to reputation. In this paper we focus on a sample data, consisting of 808 cyber attacks observed in 2017, the year in which most data was observed. Severity levels are reported according to the type and technique of attacks (which can be seen as event types), the victims and their country of origin. We remark that the considered sample data may represent a partial situation, less critical than the real one. This because many attacks may not be disclosed, or may be disclosed very late.

Here, we focus on detecting the main factors which may affect the severity degree. For this purpose, we consider two rank regression models which differ in terms of the variables taken into account. The first rank regression model is built on all the explanatory variables appearing in our dataset. Thus, cyber attacks, attack techniques, victim type and continent are introduced into the model. A second rank regression model was specified by removing from the full model the continent variable. This in order to assess if the geographical area where the cyber attacks occur may impact on the severity degree.

In Tables 1 and 2, we report the significant effects (at a significance level $\alpha = 0.05$) provided by the full and reduced rank regression models. We remark that both models are significant yielding a p -value smaller than 0.001. In addition, the full

rank regression model yields $R^2 = 0.6183$ and the reduced rank regression model yields $R^2 = 0.6176$.

Table 1: Significant effects from the fitted full rank regression model at $\alpha = 0.05$.
Categorical variable reference level: cyber attack (first block): Cybercrime; victim type (second block): Automotive; attack technique (third block): 0-day

<i>Coefficient</i>	<i>Estimate</i>	<i>p-value</i>
Intercept	87.42	0.02678
Espionage/Sabotage	-231.38	<0.001
Hactivism	-39.210	0.00663
Information warfare	-222.17	<0.001
Entertainment/News	117.14	0.03345
GDO/Retail	139.97	0.01743
Online Services/Cloud	136.11	0.01496
Research-Education	142.26	0.01057
Phishing/Social Engineering	120.27	0.01763
Unknown	99.670	0.04516

Table 2: Significant effects from the fitted reduced rank regression model (without continent variable) at $\alpha = 0.05$.

Categorical variable reference level: cyber attack: Cybercrime (first block); victim type (second block): Automotive; attack technique (third block): 0-day

<i>Coefficient</i>	<i>Estimate</i>	<i>p-value</i>
Intercept	175.65	0.01615
Espionage/Sabotage	-231.88	<0.001
Hactivism	-38.99	0.00672
Information warfare	-221.71	<0.001
Entertainment/News	115.53	0.03549
GDO/Retail	138.18	0.01855
Online Services/Cloud	135.52	0.01514
Research-Education	140.07	0.01158
Phishing/Social Engineering	120.63	0.01708
Unknown	100.21	0.04357

From Table 1 the main interesting issue that arises from the full rank regression model is the absence of the continent variable among the significant effects. This leads us to believe that such a variable may be omitted from the model since without any impact on the cyber risk. As a further consideration, note that both models provide the same variable effect sign on the severity degree.

We now move to model validation, with the purpose of selecting the model with the highest predictive accuracy, here measured by the RG metric. To have a more

exhaustive picture, we also include the computation of the RMSE, as an example of traditional validation criterion. The results are displayed in Table 3.

Table 3: RG measure for the full and reduced (without continent variable) rank regression models

Model	RG	RG_{norm}	RMSE
Full rank regression model	63.185	0.739	105.196
Reduced rank regression model (without continent variable)	63.111	0.738	105.284

From Table 3, the difference between the RG values computed in absolute terms on the two models is really tiny. This happens also for the RMSE. In addition, since $RG_{max} \simeq 85.492$, it follows that the full and reduced rank regression models explain about the 74% of the variable ordering showing that there is no relevant difference between the models. Thus, the choice falls on the model without the continent variable.

With the aim of further validate the proposed model selection measure, we led an additional analysis in which the full rank regression model is preserved with the same variables but the reduced rank regression model is built including the variables attack techniques, victim type and continent. The only variable excluded from the model is cyber attack type. Also in this case the reduced rank regression model without the cyber attack variable is significant ($p\text{-value} < 0.001$). The goodness of fit measure $R^2 = 0.4806$, which is greatly smaller than the R^2 value obtained on the full rank regression model. For the sake of brevity, we do not provide the table displaying the significant effects, but we only point out that, compared with the reduced rank regression model without the continent variable, in the reduced rank regression model without the cyber attack variable, Entertainment/News is no more significant while DDoS, malware, Malware and Vulnerabilities become significant.

Results in terms of RG and RMSE are reported in Table 4.

Table 4: RG measure for the full and reduced (without cyber attack variable) rank regression models.

Model	RG	RG_{norm}	RMSE
Full rank regression model	63.185	0.739	105.196
Reduced rank regression model (without cyber attack variable)	47.426	0.555	122.706

From Table 4, the reduced rank regression model without the cyber attack variable only explains about the 55.5% of the variable ordering showing how the role

played by the cyber attack cannot be neglected since the loss in terms of severity explanation is strongly evident. The same can be said if referring to the RMSE.

4 Conclusions

In this paper we have discussed a novel model to measure cyber risks, which takes the ordinal nature of the disclosed data correctly into account. The proposed model can be employed as a simple and effective measurement to prioritise cyber risk, as shown in our case-study. Its application to a real cyber loss database, measured at the ordinal level, reveals that the proposed tools are indeed able to detect the main factors affecting the cyber risks.

References

1. Agresti, A.: Analysis of ordinal categorical data, Second Edition, Wiley, New York (2010)
2. Clusit: 2018 Report on ICT security in Italy (2018)
3. Cox, L.A.Jr: Evaluating and improving risk formulas for allocating limited budgets to expensive risk-reduction opportunities. *Risk. Anal.*, **32**, 1244–1252 (2012)
4. Edgar, T.W., Manz, D.O.: Research Methods for Cyber Security, Elsevier (2017)
5. Giudici, P., Raffinetti, E.: A Rank Graduation measure to assess predictive accuracy. Technical report, submitted (2019)
6. Lorenz, M.O.: Methods of Measuring the Concentration of Wealth. *J. Am. Stat. Assoc.* **9(70)**, 209–219 (1905)
7. McCullagh, P.: Regression Models for Ordinal Data. *J. Roy. Stat. Soc. B Met.* **42(2)**, 109–142 (1980)

Vector Error Correction models to measure connectedness of bitcoin exchange markets

Modelli di Vector Error Correction per misurare la connessione delle piattaforme di scambio di bitcoin

Paolo Giudici and Paolo Pagnottoni

Abstract Bitcoins are traded on various exchange platforms and, therefore, prices may differ across trading venues. We aim to investigate return connectedness across eight of the major exchanges of Bitcoin, both from a static and a dynamic viewpoint. To this end, we extend the order-invariant forecast error variance decomposition proposed by Diebold and Yilmaz (2012) to a generalized vector error correction framework. Our results suggest that there is strong connectedness among the exchanges, as expected, although some of them behave dissimilarly. We identify Bitfinex and Coinbase as leading exchanges during the considered period, while Kraken as a follower exchange. We also obtain that connectedness across exchanges is strongly dynamic, as it evolves over time.

Abstract *I bitcoin sono scambiati su varie piattaforme e, pertanto, i prezzi possono differire tra le piattaforme di trading. Il nostro obiettivo è quello di investigare la connessione nei rendimenti di otto delle maggiori piattaforme di scambio di Bitcoin, sia dal punto di vista statico che da quello dinamico. Pertanto, estendiamo la tecnica di decomposizione della varianza dell'errore di previsione invariante all'ordine di Diebold e Yilmaz (2012) al contesto di modelli Vector Error Correction generalizzati. I nostri risultati suggeriscono che c'è una forte connessione tra gli exchange, come ci si aspettava, nonostante alcuni si comportino in modo dissimile. Identifichiamo Bitfinex e Coinbase come exchange leader durante il periodo considerato, laddove Kraken come follower. Otteniamo inoltre che la connessione tra gli exchange è fortemente dinamica, ed evolve con il tempo.*

Key words: Bitcoin; Market risk; Market linkages; Vector autoregression; Vector error correction; Forecast error variance decomposition; Spillovers

Paolo Giudici

University of Pavia, Via S.Felice 5, 27100, Pavia (PV), e-mail: paolo.giudici@unipv.it

Paolo Pagnottoni

University of Pavia, Via S.Felice 5, 27100, Pavia (PV), e-mail: paolo.pagnottoni01@universitadipavia.it

1 Introduction

Connectedness is gaining much importance in financial econometrics and risk management. The study of return connectedness is key to assess market risk and, in particular, to understand which are the market exchanges whose shocks in price are transmitted to the others; or which are those that receive shocks from the others and adjust their price consequently. In other words, the study of connectedness across market exchanges is fundamental for price discovery purposes, that is, to determine the leader-follower relationships between markets.

The available literature on price discovery for cryptocurrency exchanges is quite limited. The first researchers addressing this issue are Brandvold *et al.* (2015), who found that Mt.Gox - a leading exchange which went bankrupt right after their analysis - and BTC-e were the leaders of price discovery. A more recent study by Pagnottoni *et al.* (2018) makes use of the Hasbrouck (1995) and Gonzalo & Granger (1995) methodologies taking into account also the impact of exchange rates. They find Chinese exchanges to be the ones leading price discovery during their analyzed period. More recently, Giudici & Abu-Hashish (2018) employ a VAR model, that embeds into its correlation structure the connectedness among eight Bitcoin exchanges. All previous papers have the merit of being the first ones in the field but, on the other hand, they are limited as they lack either a modelling strategy that can make them predictive (as is the case of the first two papers) or do not take into account important econometric aspects such as cointegration and stationarity of the considered series (as is the case with the third paper).

We aim to improve the latter contributions, and suggest a model that, while fully grounded on an econometric approach, builds a comprehensive statistical model. To this aim, we rely on the order-invariant forecast error variance decomposition proposed by Diebold & Yilmaz (2012). This is line with the recent paper by Corbet *et al.* (2018), who investigates the dynamic relationships, particularly volatility spillovers, between major cryptocurrencies (Bitcoin, Litecoin and Ripple) and other financial assets through the Diebold & Yilmaz (2012) generalized variance decomposition technique.

A limitation of Corbet *et al.* (2018) is that they employ a generalized VAR model, which does not take cointegration of the series into account, a phenomena that is particularly evident when studying connectedness among market exchanges prices concerning the same asset, such as the Bitcoin.

Indeed, Diebold & Yilmaz (2012) develop measures of directional spillovers in a generalized VAR framework, which are suitable for several applications. However, when the time series under consideration are integrated of order one ($I(1)$), the VAR model is not suitable to model them, but it may be fine to model their first difference it may be the case. However, if the same time series additionally show a significant co-movement around a common stochastic trend, i.e. they are cointegrated, Engle & Granger (1987) show it is reasonable to model them as a VEC model (VECM), whose error correction term takes into account the common stochastic trend driving prices.

In line with the previous comment, here we extend Diebold & Yilmaz (2012) with a generalized VECM. This allow us to study connectedness at different levels: pairwise and system-wise, as well as both from a static and time-varying point of view, accounting for the common stochastic trend driving the fundamental Bitcoin price.

We therefore contribute, from a methodological viewpoint, to the econometric literature, proposing an extension of the (Diebold & Yilmaz, 2012) connectedness measure, which relies on VECM rather than VAR models. The model allows to shed further light on price discovery in Bitcoin markets, extending the conclusions in Corbet *et al.* (2018) and Giudici & Abu-Hashish (2018) and, in particular, characterizing which are the leaders and followers in price formation among the considered exchanges, along time.

2 Methodology

We denote the Bitcoin price of an exchange i at time t as X_t^i , while its logarithm as x_t^i . Given the dynamics of Bitcoin prices over time, it is reasonable to expect that in our case we deal with non-stationary time series - arguably $I(1)$. Moreover, we are in the situation in which the same asset is traded across different platforms. In such a framework the law of one price prescribes that prices related to the same good should not deviate in the long run. Strictly speaking, the no-arbitrage condition implies, when Bitcoin prices are expressed in the same currency, that there exist linear combination of their (log-) prices yielding a stationary process.

The considerations from above make us expect there is a cointegration structure among our variables. Thus, the theoretical econometric framework delineated by Engle & Granger (1987) suggests us to exploit the VECM.

Note that continuous returns are the first difference of the exchange prices in log levels. Collecting those first differences into the vector $\Delta x_t = (\Delta x_t^1, \dots, \Delta x_t^i, \dots, \Delta x_t^n)'$ with $i = 1, 2, \dots, n$ exchanges, the model assumes the following form:

$$\Delta x_t = \alpha \beta' x_{t-1} + \sum_{i=1}^{k-1} \zeta_i \Delta x_{t-i} + \varepsilon_t \quad (1)$$

with α being the $(n \times h)$ adjustment coefficient matrix, β the $(n \times h)$ cointegrating matrix, ζ_i the $(n \times n)$ parameter matrices with $i = 1, \dots, n$, k the autoregressive order and ε_t is a zero-mean white noise process having variance-covariance matrix Σ . We denote as h the cointegrating rank. In our case, the time series in levels should show one common stochastic trend, i.e. economic theory suggests that the cointegrating rank of the system is $h = n - 1$.

In our paper we rely, as in Diebold & Yilmaz (2012), on the Koop *et al.* (1996) Pesaran & Shin (1998) (KPPS) H-step-ahead forecast errors. They have the

advantage to be invariant to the variable ordering, unlike the popular although restrictive Cholesky factorization, which would require an ordering of the Bitcoin exchange prices *a priori* with regards to the influence of shocks across the system variables.

Taking two generic variables x_i and x_j , Diebold & Yilmaz (2012) define the own variance shares as the proportion of the H -step ahead error variance in forecasting x_i due to shocks in x_i itself, $\forall i = 1, \dots, n$, whereas the cross variance shares (spillovers) are defined as the H -step ahead error variance in predicting x_i due to shocks in x_j , $\forall i = 1, \dots, n$ with $j \neq i$. That said, using $\theta_{ij}^g(H)$ to denote the KPSS H -step forecast error variance decompositions, with $H = 1, \dots, n$, we have:

$$\theta_{ij}^g(H) = \frac{\sigma_{jj}^{-1} \sum_{h=0}^{H-1} (e_i' \Psi_h \Sigma e_j)^2}{\sum_{h=0}^{H-1} (e_i' \Psi_h \Sigma \Psi_h' e_j)} \quad (2)$$

where σ_{jj} is the standard deviation of the innovation for equation j and e_i represents the selection vector with one as element i and zeros elsewhere.

From the use of the total contributions to the forecast error variance decomposition we estimate the Total Spillover Index (TSI), the Directional Spillover Indexes (DSI) and the Pairwise Net Spillover Indexes (PNSI).

By means of the variance decompositions, these measures allow to analyze exchange spillovers both from a system-wide and a net pairwise viewpoint. Outcomes are presented in the results Section.

3 Data

We consider for our empirical analysis what is arguably the most relevant cryptocurrency nowadays existing: Bitcoin. We indeed examine Bitcoin exchange prices denominated in USD on a daily basis during a time-frame from 18 May 2016 and 30 April 2018, as in Giudici & Abu-Hashish (2018). With the aim to study system-wise connectedness as well as the pairwise one among Bitcoin trading venues, we consider eight Bitcoin exchanges, i.e. Bitfinex, Coinbase, Bitstamp, Hitbtc, Gemini, ItBit, Kraken, Bittrex. We remark that the investigated exchanges are geographically widespread, with their daily trading volumes summing up to more than 75% of the total.

4 Results

For the sake of brevity of this manuscript, we will present only dynamic results. In detail, we fix a rolling estimation window of 125 days and a 10-step ahead forecast horizon for the variance decomposition. As already anticipated, the dynamic

analysis should be able to explain the outcomes of the "average" connectedness measures derived before. Again for the sake of brevity, we decided to confine this manuscript's analysis to the Net Spillover Indexes (NSI), as illustrated in Figure 1.

What we are interested in is the dynamic leader-follower relationship among exchanges. In this regard, the results are again in line with those obtained in the full sample analysis: besides Bittrex, Bitfinex appears to be the exchange receiving the least and contributing the most to others in terms of return spillovers over time, immediately followed by Coinbase. In both cases, the magnitude of their influence varies over time. Notice also that Kraken is the exchange being the most influenced from others. From the beginning of May 2017, its return spillover contribution to others starts declining, whereas the one transmitted by others begins to rise. It is interesting to notice that Kraken's follower behaviour begins with the surge in Bitcoin prices, a day in which exchanges connectedness are arguably expected to experience some changes. This marks the beginning of a "follower phase" for Kraken, which lasts until the end of the sample, where we see its net contribution converging to its previous values.

5 Conclusion

In the paper we have proposed a novel approach to estimate return spillovers across Bitcoin exchanges, that relies on a cointegrated vector error correction framework.

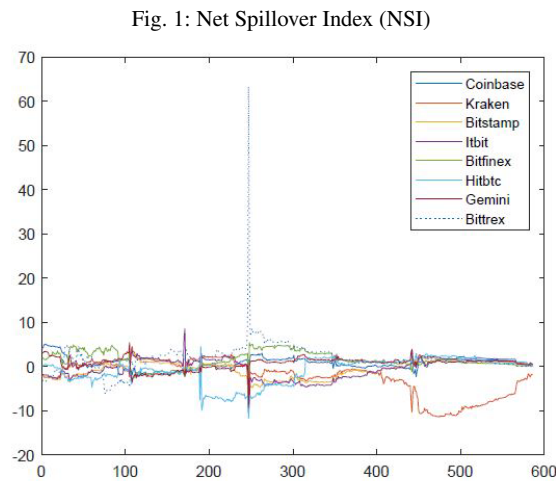


Fig. 2: *Note:* The figure from above illustrates the dynamic net return spillover indexes for the period 22 September 2016 - 30 April 2018. The rolling window set for the estimations is $w = 125$ days. Values are expressed in percentage terms.

We believe that our proposal can be extended, without loss of generality, to other cryptocurrencies, as well as to traditional markets.

From a methodological point of view, we extend the generalized variance decomposition technique introduced by Diebold & Yilmaz (2012). While Diebold & Yilmaz (2012) derive measures for the directional spillovers across markets within a generalized VAR framework, we adapt their methodology to the case in which the same asset (Bitcoin) is traded on multiple exchanges. Since we deal with $I(1)$ price series related to a unique asset - i.e. arguably cointegrated -, we rely on a generalized VECM framework to derive directional spillovers.

From an empirical viewpoint, our results show that, Bitfinex and Coinbase are the leader exchanges in the price formation process, transmitting a significant portion of return spillover to other exchanges. Moreover, we find Kraken among the follower exchanges. This arguably confirms that the exchanges in which most of the trading volumes lie are generally also the ones giving substantial contribution to other markets from a price discovery point of view.

In addition, our results show that return spillovers across Bitcoin exchanges are dynamic and sensibly evolve over time. In other words, the leader-follower compositions are not constant in time and may consistently evolve. This is in line with what observed by Brandvold *et al.* (2015) and Pagnottoni *et al.* (2018), who concluded that information shares are dynamic and may consistently evolve over time.

References

- BRANDVOLD, M., MOLNÁR, P., VAGSTAD, K., VALSTAD, A., & OLE, C. 2015. Price discovery on Bitcoin exchanges. *Journal of International Financial Markets, Institutions and Money*, **36**(C), 18–35.
- CORBET, S., MEEGAN, A., LARKIN, C., LUCEY, B., & YAROVAYA, L. 2018. Exploring the dynamic relationships between cryptocurrencies and other financial assets. *Economics Letters*, **165**, to appear.
- DIEBOLD, FRANCIS X., & YILMAZ, KAMIL. 2012. Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of Forecasting*, **28**(1), 57–66.
- ENGLE, ROBERT F., & GRANGER, CLIVE WJ. 1987. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251–276.
- GIUDICI, P., & ABU-HASHISH, I. 2018. What determines prices in bitcoin markets: a network VAR approach. *Finance Research Letters*.
- GONZALO, J., & GRANGER, C. 1995. Estimation of Common Long-Memory Components in Cointegrated Systems. *Journal of Business & Economic Statistics*, **13**(1), 27–35.
- HASBROUCK, J. 1995. One Security, Many Markets: Determining the Contributions to Price Discovery. *The Journal of Finance*, **50**(4), 1175–1199.
- KOOP, GARY, PESARAN, M HASHEM, & POTTER, SIMON M. 1996. Impulse response analysis in nonlinear multivariate models. *Journal of econometrics*, **74**(1), 119–147.
- PAGNOTTONI, P., DIRK, B., & DIMPFL, T. 2018. Price Discovery on Bitcoin Markets. *Working Paper*.

Vector Error Correction models to measure connectedness of bitcoin exchange markets

PESARAN, M HASHEM, & SHIN, YONGCHEOL. 1998. An autoregressive distributed-lag modelling approach to cointegration analysis. *Econometric Society Monographs*, **31**, 371–413.

Estimation of lineup efficiency effects in Basketball using play-by-play data

L'uso dei dati del play-by-play per la stima degli effetti di quintetto nella pallacanestro

Luca Grassetti, Ruggero Bellio, Giovanni Fonseca and Paolo Vidoni

Abstract The paper aims at defining a data-driven approach to team management in basketball. A model-based strategy, based on a modification of the adjusted plus-minus approach, is proposed for the analyses of the match progress. The main idea is to define a model based on the 5-man lineups instead of the single players. In this framework, given the large number of possible lineups, the regularization issue is quite relevant. The empirical application is based on the data of the current Italian championship (Serie A1). The play-by-play data are considered along with some information resulting from the game box scores.

Abstract *Questo contributo propone un approccio “data driven” al processo di gestione delle squadre nella pallacanestro. In particolare, si sviluppa una strategia “model-based” per l’analisi della progressione delle partite, basata su una versione modificata del modello per l’adjusted plus-minus. L’obiettivo è definire il modello sulla base dei quintetti anziché dei singoli giocatori. In questo contesto, data l’elevata numerosità dei possibili quintetti, il ricorso ai metodi di regolarizzazione risulta essere molto rilevante. L’analisi empirica è basata sui dati dell’attuale campionato della Lega Basket Italiana (Serie A1). Il play-by-play delle partite è considerato congiuntamente ad alcuni dati raccolti dai box score delle partite.*

Key words: Basketball Analytics, Data-driven decision process, Play-by-Play data, Statistical Model, Web-crawling.

Luca Grassetti

Department of Economics and Statistics, University of Udine, Italy
Via Tomadini, 30/A - 33100 Udine (UD) e-mail: luca.grassetti@uniud.it

Ruggero Bellio

e-mail: ruggero.bellio@uniud.it

Giovanni Fonseca

e-mail: giovanni.fonseca@uniud.it

Paolo Vidoni

e-mail: paolo.vidoni@uniud.it

1 Introduction

Sports analytics for team games has principally two main purposes: match outcome prediction and analysis of performance. Even if the two issues are related, methods adopted may be quite different. Basketball makes no exception and outcome prediction literature is based on some model assumption on the final outcome. For example, [16] and [17] assume Poisson distribution for each team final match score in order to compute match-winning probabilities, as usually done in soccer literature. Further extensions are provided by [8] that uses logistic regression in a Markov Chain setting and by [11] that introduces a time variation for the strength of teams in the model. On the other side, there are results based on classification and automatic algorithms, as in [20] and [21] where the authors adopt a machine learning approach in order to predict match results.

Players performance analyses are usually based on box score statistics that are widely available for basketball matches. Moreover, nowadays, data are collected on a real-time basis during the game and possession outcomes are used for describing a single player's performance. For example, [2] relates the individual player contribution to the match-winning probability of the team at different game moments.

A remarkable exception is the estimation of player performances based on the so-called adjusted plus-minus (APM) method. First introduced in an influential contribution by [15], the computation of the APM is based on play-by-play data aggregated in *shifts*. A shift is defined as a period of playing time without any substitution for either team. The APM is computed by considering a linear regression model for the point differential during each shift as the response variable, with regressors given by signed dummy variables for every single player involved in the shift. The estimation is carried out from the perspective of the home team, with the consequential definition of the point difference and the sign of player dummy variables.

The APM method has a strong appeal since the estimated coefficients can be interpreted as (net) player efficiency measures, i.e. they are adjusted for the other players on the field. Not surprisingly, the measure was readily adopted by NBA data analysts. At the same time, the original method is prone to difficulties, since it entails sparse design matrices and multicollinearity. In other words, it is a typical setting where some form of regularization is called for, and, indeed, the Regularized APM (RAPM) was proposed in [18]. The RAPM method employs ridge regression for the estimation of player efficiency and, as summarised in [4], it is more accurate, robust and stable than the original APM. The method was adopted also for the analysis of the players of the Major League Soccer [7] and the National Hockey League (see [9] and [10]).

The proposal of this paper builds upon the RAPM setting, with three important variations. The main innovation is that instead of focusing on the performances of individual players, we focus on the performances of the entire lineups (5-man units) on the field, with the idea that player performances may depend on the interaction with teammates and on the contrastive action of the lineup of the opposite team on the field. From a statistical perspective, regularization is even more essential for the estimation of lineup efficiency than for that of individual players, so that

we may consider also other approaches next to ridge regression, such as empirical Bayes and boosting (see, for instance, [3]). A final distinctive feature of the current proposal is the adoption of a performance index rating as the response variable, thus considering a more comprehensive measure than point differential. The overall aim is to provide a useful tool for coaches that highlights the strength and the weakness of the different available lineups, offering some added value with respect to the evaluation of the performance of individual players.

The paper is organized as follows. Section 2 illustrates the data used for the analysis, obtained from the Italian Basketball League. Since the data wrangling process presents some features that may be of interest for statisticians, it is illustrated with some details. Section 3 introduces the model adopted for the estimation of lineup efficiency, and it reports some results for the case study of interest. Finally, Section 4 contains a brief discussion and some concluding remarks.

2 Data wrangling and data exploration

The empirical analysis described in this paper is based on a dataset regarding the Italian Basketball League (Serie A1). In particular, the matches of the first round of the current championship 2018/2019 are considered. The dataset collects the play-by-play information along with the matches box scores, which are made available by the league website (www.legabasket.it). The plays are then aggregated in shifts, which is the aggregation level considered in the statistical analysis. In this section, the data wrangling process is described, emphasizing some key aspects which could be interesting from the statistical perspective, and the results of a preliminary data exploration are briefly presented.

2.1 Data collection

In order to collect the data from the Italian Basketball League website, we use the R statistical software [13]. In particular, for performing the data scraping process we consider the `rvest` package [19]. The `scrapeR` [1] and `Rcrawler` [6] packages may also be used to this end. For every single match, we collect both the box scores data and the play-by-play information. A play is defined as an event during the possession involving a positive or negative evaluation (see Table 1).

For the collection of the box scores data of a single match, the associated specific web page is parsed, the tables are then identified in the text using the function `html_nodes` and, finally, the box scores information are organized in a data frame with the function `html_table`.

The play-by-play data can be obtained using a similar procedure, but some useful tricks have to be considered in order to produce a ready-to-use dataset. To this end, the information available in the play-by-play table is collected as raw text extracted

using the command `html_text`. Furthermore, the textual output is pre-processed using the `stringi` [5] package, with the aim to code the information in a usable way. Each play is finally recorded by considering the following features:

- player finalizing the play,
- possible intermediate events in a play (substitutions, time-outs and so on),
- outcome of the play (classified according to different potential categories),
- quarter,
- minute in the quarter,
- home and the away teams,
- team of the finalizing player.

These data are then completed using the box scores tables. In particular, by considering the information on the starting five, we are able to reconstruct the 5-man unit involved in each play. In fact, whenever the event in the play-by-play dataset is a substitution, the change in the 5-man unit is recorded. This piece of information is crucial in order to aggregate the plays in shifts, as required for subsequent analyses.

2.2 Data cleansing

The data cleansing process is a further important step in order to define the dataset in the required form. At first, a check on the names of the involved players is required, since they are sometimes reported with errors. Furthermore, the players having an average play time shorter than five minutes are removed from the analysis, so that the 5-man units involving these players will present one or more anonymized individuals, as proposed in the APM literature (see for example [4]). We call these individuals *dummy players*. The original lineups are also recorded.

A numerical variable is then defined by some specific outcomes produced in the plays. In particular, the scores reported in Table 1 are assigned to the offensive team, and opposite scores are assigned to the defensive team. The scores are defined by considering only those events deemed as the most relevant for the outcome of the play.

Table 1 Scores of the events used in the computation of the outcome measure for each play.

Value	Events
-1	missed free-throw, turnover or offensive foul
-0.5	missed shot (2 points or three points shots)
0.5	assist
1	steal, offensive or defensive rebound, block, scored free-throw or received foul
2	scored shot
3	scored three-pointer

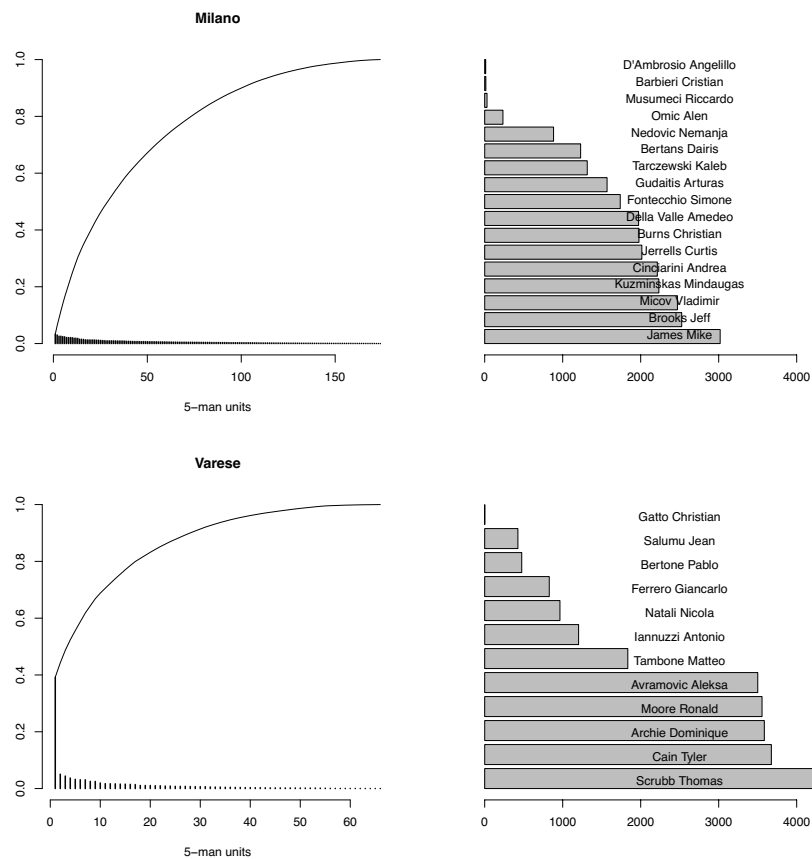
The records regarding turnovers caused by steals and personal defensive foul corresponding to a received foul are removed in order to prevent duplicated infor-

mation. Assists could also be removed from the analysis if they are deemed as not particularly relevant. The sum of the above single scores produces a result which is not far from the final result of the match, and it has the merit of accounting for some non-scoring key events.

2.3 Some summary statistics

The data used for the analysis concern 120 matches, resulting in 4108 shifts with more than 41000 plays, corresponding to around 21000 possessions. The total number of lineups is around 2000. The latter number would be actually slightly larger had we not replaced each player with low playing time per game by a single team-specific dummy player.

Fig. 1 Distribution of the number of plays for Milano and Varese team.



The left panels of Figure 1 reports the distribution of the number of plays for lineups for two selected teams, Milano and Varese, that are the two teams with the largest and smallest number of lineups, respectively. The right panels of the same figure report the total number of plays for the players with the highest totals.

3 Estimation of lineup effects

The starting point for the estimation of lineup effects is a simple model for the score of the t shift, with $t = 1, \dots, T$,

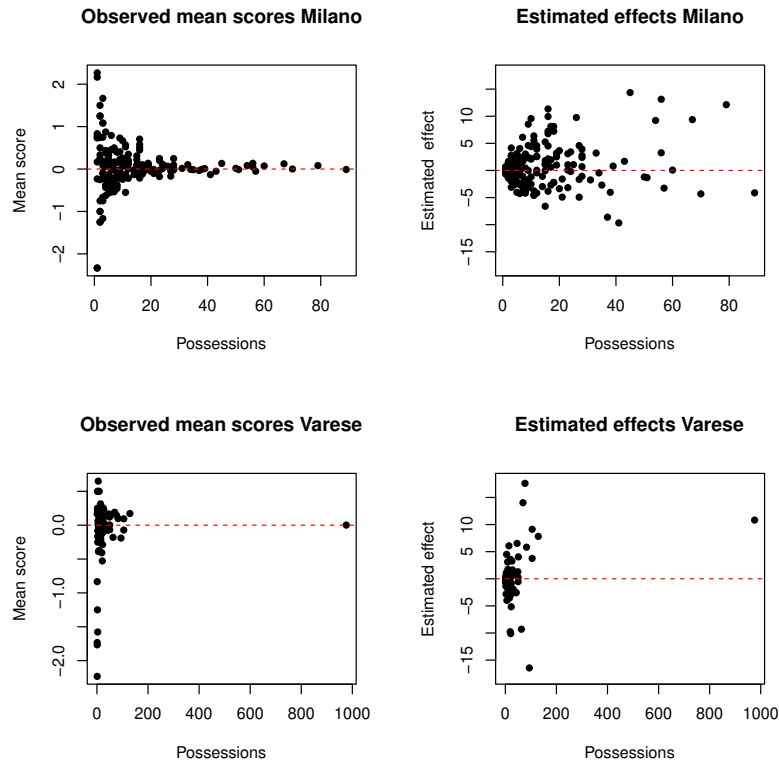
$$y_t = \beta_0 + \mu_{h[t]} - \mu_{a[t]} + \varepsilon_t. \quad (1)$$

Here we consider the entire data set encompassing all the available matches so that $T=4108$. The response variable in (1) is given by the difference between the mean outcome for the home team and the mean outcome of the away time for each shift, where the mean is over the number of possessions for each team. In case only one team is involved in a given shift, the mean outcome of the opposite team is replaced by the grand mean over the entire sample, along the lines of what done in the APM literature. The notation $h[t]$ and $a[t]$ identifies the lineup for the home and away team for shift t , respectively. More precisely, $h[t]$ and $a[t]$ assume a value in the set $1, \dots, N$, where N is the total number of lineups in the dataset ($N = 1998$ in the application). The model specification could be made more complex by including some covariate effects, but for the sake of illustration, we focus here on a basic form.

The estimation of the lineup effects μ is based on regularized weighted linear regression, with weights corresponding to the total number of possessions of every shift. Here we report some results corresponding to the estimation based on an empirical Bayes approach, assuming normal lineup effects. The estimation has been carried out by means of the `hglm` R package [14], which allows for the inclusion of observation weights and estimates the regularization parameter by REML. Note that the results obtained with ridge regression were essentially equivalent to those obtained by empirical Bayes, but the latter approach seems preferable since it allows for straightforward inclusion of further regression covariates. Instead, the results based on boosting were somewhat unsatisfactory, exhibiting an apparent lack of shrinkage and requiring much longer computations than the two other methods.

The left panels of Figure 2 report the mean response for each lineup of the same two teams of Figure 1 against the number of possessions. The right panels of Figure 2 report instead the estimated effects $\hat{\mu}$ for the lineups of the two teams against the number of possessions. The effects are on the 100-possession scale, as customary in the APM literature. It is apparent that the estimated lineup effects adjust for the quality of the other lineups on the field. A remarkable instance is the lineup with the highest number of possessions for the Varese team, for which the apparent null observed mean score corresponds to a positive estimated effect. Indeed, this lineup has started most of the team matches, thus playing against higher-quality lineups.

Fig. 2 Observed mean scores against number of possessions (left), and lineup estimated effects against number of possessions (right), for Milano and Varese team.



4 Conclusion and ongoing research

Play-by-play data represent an invaluable source of information for the statistical analysis of basketball results. This paper has illustrated how to obtain the data and perform some pre-processing using R software tools, which may be familiar to many statisticians. The analysis reported is an extension of the RAPM approach to the analysis of lineup effects, with the important difference of considering a more comprehensive outcome measure.

Estimates of lineup effects should be of interest for team managers and data analysts, since they provide some insight about the team lineup strategy adopted. The quantification of lineup performances extends the information provided by player-based RAPM analysis. Ongoing research concerns the study of the connection existing between the two approaches, as well as the possibility of disentangling both sets of measures into more than one dimension corresponding to different aspects of the game, thus exploiting even further the availability of play-by-play data.

Acknowledgments

We are grateful to the Italian Basketball League for the permission of using the play-by-play data. This research is partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS_003.

References

1. Acton, R.M.: scrapeR: Tools for scraping data from HTML and XML documents. R package version 0.1.6. <https://CRAN.R-project.org/package=scrapeR> (2010)
2. Deshpande, S.K. and Jensen, S.T.: Estimating an NBA player's impact on his team's chances of winning. *J. Quant. Anal. Sports*, **12**, 51–72 (2016)
3. Efron, B. and Hastie, T.: *Computer Age Statistical Inference*. Cambridge University Press, Cambridge (2016)
4. Engelmann, J.: Possession-based player performance analysis in basketball (adjusted +/- and related concepts). In: *Handbook of Statistical Methods and Analyses in Sports*, pp. 231–244. Chapman and Hall/CRC (2017)
5. Gagolewski M. *et al*: R package stringi: character string processing facilities. <http://www.gagolewski.com/software/stringi/> (2019)
6. Khalil, S.: Rcrawler: web crawler and scraper. R package version 0.1.9-1. <https://CRAN.R-project.org/package=Rcrawler> (2018)
7. Kharrrat, T., Pena, J.L. and McHale, I.: Plus-minus player ratings for soccer. arXiv preprint arXiv:1706.04943, (2017)
8. Kvam, P. and Sokol, J.S.: A logistic regression/Markov chain model for NCAA basketball. *Naval Research Logistics* (2006)
9. Macdonald, B.: A regression-based adjusted plus-minus statistic for NHL players. *Journal of Quantitative Analysis in Sports* **7.3**, (2011)
10. Macdonald, B.: Adjusted plus-minus for NHL players using ridge regression with goals, shots, fenwick, and corsi. *Journal of Quantitative Analysis in Sports* **8.3** (2012)
11. Manner, H.: Modeling and forecasting the outcomes of NBA basketball games. *Journal of Quantitative Analysis in Sports*, **12**, 31–41 (2016)
12. Omidiran, D.: A new look at adjusted plus/minus for basketball analysis. MIT Sloan Sports Analytics Conference [online], (2011)
13. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
14. Ronnegard, L., Shen, X. and Alam, M.: hglm: A package for fitting hierarchical generalized linear models. *The R Journal* **2**: 20–28 (2010)
15. Rosenbaum, D.: Measuring how NBA players help their teams win. Retrieved from <http://www.82games.com/comm30.htm> (2004)
16. Ruiz, F.J.R. and Perez-Cruz, F.: A generative model for predicting outcomes in college basketball. *J. Quant. Anal. Sports*, **11**, 39–52 (2015)
17. Shen, K.: Data analysis of basketball game performance based on bivariate poisson regression model. *Computer Modelling & New Technologies*, **18**, 474–479 (2014)
18. Sill, J.: Improved NBA adjusted +/- using regularization and out-of-sample testing. In: *Proceedings of the 2010 MIT Sloan Sports Analytics Conference* (2010)
19. Wickham H.: rvest: easily harvest (scrape) web pages. R package version 0.3.2. <https://CRAN.R-project.org/package=rvest> (2016)
20. Yang, J.B. and Lu, C.-H.: Predicting NBA Championship by learning from history data. *Proceedings of Artificial Intelligence and Machine Learning for Engineering Design* (2012)
21. Zimmermann, A., Moorthy, S. and Shi, Z.: Predicting NCAAAB match outcomes using ML techniques - Some results and lessons learned. *Proceedings ECML 2013*, 69–78 (2013)

Trajectory clustering using adaptive squared distances

Clustering di triettorie attraverso distanze adattative quadratiche

Antonio Irpino

Abstract The paper deals with the clustering of trajectories of moving objects. A k-means-like algorithm based on a Euclidean distance between piece-wise linear curves is used. The main novelty of the paper is the opportunity of considering in the clustering procedure a step that automatically weights the importance of sub-trajectories of the original ones. The algorithm uses an adaptive distances approach and a cluster-wise weighting. The proposed algorithm is tested against some work-bench trajectory datasets.

Abstract *Abstract in Italian*

Key words: K-means, Adaptive Distances, Trajectories

1 Introduction

Nowadays, surveillance systems or the ubiquity of global positioning system (GPS) sensors integrated into devices produce a huge amount of data about moving objects in terms of trajectories. The extraction of patterns from trajectory data is increasingly challenging and demanding. Clustering is a very useful tool for extracting regularities from data and trajectory clustering is gaining a great interest in the data analysis community for its peculiarities.

Thus, the problem of clustering trajectories depends on how trajectories are compared, or if a trajectory is considered as a set of sub-trajectories or not. Indeed, in the latter case, the aim is to extract sub-patterns that represents, at least, interesting parts of the trajectories. On the other hand, considering a trajectory as a unicum involves to find groups of similar trajectories.

Antonio Irpino

Università degli studi della Campania “L. Vanvitelli”, Dipartimento di Matematica e Fisica, Viale A. Lincoln 5, Caserta, Italy, e-mail: antonio.irpino@unicampania.it

Trajectories clustering looks for groups of trajectories, or of sub-trajectories, such that they represents a movement pattern in the data. The subject is surveyed in [5]. A similar approach is used for monitoring animal behaviour [1]. In this case, probabilistic model are assumed as valid and the extraction of patterns depends on the model choice. In the literature, two main algorithms are considered: the [6] and the [4]. In [6], a distance between sub-trajectories is defined and the algorithm implements an extension of a density based algorithm for grouping set of sub-trajectories. In [4], the idea is to estimate k predefined vector fields that represent group of trajectories observed in a 2D space. This application, is inspired by the problem of monitoring and predicting storm or hurricane paths. In a functional data analysis approach, a trajectory is considered as a curve in a 2D or 3D space. In order to be analysed a smoothing, interpolation, or alignment step is performed and then the trajectory are analysed [8].

In this paper, we consider a k-means approach for grouping trajectories. We show how to decompose the Euclidean distance between two trajectories. We use such a decomposition for defining some aspects of the compared trajectories and we enrich the k-means algorithm with a step that automatically assign a relevance weights to the aspects.

Finally, we show some preliminary results on some benchmark datasets.

2 Data and distances

A trajectory is a sequence of ordered space-time points (namely, a point has two or three spatial coordinates and a time-stamp), where the order follows time. A trajectory P_i is a collection of ordered pairs of data (s_j^i, t_j^i) , $j = 1, \dots, T$, sampled in T time-points where s_j^i is a spatial location (namely, a 2D or a 3D vector of spatial coordinates) and t_j^i is a time-stamp. A trajectory can be enriched with other data recorded at each time-point but we do not consider it in this paper. Considering the order provided by the time-stamps, we can consider a trajectory P as a curve represented in a 2D (or 3D space).

A set of N trajectories is a collection of trajectories denoted as P_i . We assume that each trajectory may have a different number of sampled time-points T_i . Clustering is based on a distance/dissimilarity measure between objects. In our case, the computation of a distance between two trajectories may require a normalization step for comparing them. Such a step, depending on the application domain and on the aim of analysis, may be questionable.

A first problem with trajectories is that they may have different time length. We assume that all the trajectories can be *normalized* such that $\tau_0^i = 0$ and $\tau_{T_i}^i = 1$. This is obtained by the following formula:

$$\tau_j^i = \frac{t_j^i - t_0^i}{t_{T_i}^i - t_0^i} \quad (1)$$

Further, we assume that between two consecutive points, the path is a segment starting from \mathbf{s}_j^i and ending at \mathbf{s}_{j+1}^i . These two assumptions should be justified by the problem at hand. For example, if we are monitoring a scene in a bounded space where people can flow from a point to another, or cars traveling on a route, this assumption is quite reasonable. The assumption could be considered too strong when trajectories have very different time length, that is the case of monitoring series of tracks of hurricanes or tropical storms in the Atlantic Ocean.

The hypothesis that a trajectory is a piece-wise linear curve, is computationally useful for computing a continuous version of the Euclidean distance between two trajectories.

Under this assumption, we can consider the Euclidean distance between two 2D trajectories¹ having the same k time stamps normalized in $[0, 1]$ as follows. Given two normalized trajectories $P_1 = \left\{ \{(x_0^1, y_0^1), 0\}, \dots, \{(x_j^1, y_j^1), \tau_j^1\}, \dots, (x_{T_1}^1, y_{T_1}^1), 1 \} \right\}$ and $P_2 = \left\{ \{(x_0^2, y_0^2), 0\}, \dots, \{(x_j^2, y_j^2), \tau_j^2\}, \dots, \{(x_{T_2}^2, y_{T_2}^2), 1\} \right\}$. Considering the piece-wise linear assumption, and constant speed between each pair of sampled points, it is possible to express the two trajectories with a common set of τ 's by a linear interpolation. Once the two trajectories are registered such that they have the same normalized $L \in [\min(T_1, T_2), (T_1 + T_2)]$ time-stamps we compute the squared Euclidean distance between P_1 and P_2 as follows:

$$\begin{aligned} d_E^2(P_1, P_2) &= \int_0^1 \left[(x_1(\tau) - x_2(\tau))^2 + (y_1(\tau) - y_2(\tau))^2 \right] d\tau = \\ &= \sum_{\ell=1}^L (\tau_\ell - \tau_{\ell-1}) \left\{ |\bar{x}_1(\ell) - \bar{x}_2(\ell)|^2 + |\bar{y}_1(\ell) - \bar{y}_2(\ell)|^2 + \right. \\ &\quad \left. + \frac{1}{3} \left[|\dot{x}_1(\ell) - \dot{x}_2(\ell)|^2 + |\dot{y}_1(\ell) - \dot{y}_2(\ell)|^2 \right] \right\} \end{aligned} \quad (2)$$

where:

- $\bar{x}_1(\ell) = \frac{x_1(\tau_\ell) + x_1(\tau_{\ell-1})}{2}$, $\bar{x}_2(\ell) = \frac{x_2(\tau_\ell) + x_2(\tau_{\ell-1})}{2}$, $\bar{y}_1(\ell) = \frac{y_1(\tau_\ell) + y_1(\tau_{\ell-1})}{2}$, and $\bar{y}_2(\ell) = \frac{y_2(\tau_\ell) + y_2(\tau_{\ell-1})}{2}$. The points $(\bar{x}_1(\ell), \bar{y}_1(\ell))$ and $(\bar{x}_2(\ell), \bar{y}_2(\ell))$ are, respectively, the centers of the segment that starts from $(x_1(\tau_{\ell-1}), y_1(\tau_{\ell-1}))$ and arrives at $(x_1(\tau_\ell), y_1(\tau_\ell))$, respectively, the centers of the segment that starts from $(x_2(\tau_{\ell-1}), y_2(\tau_{\ell-1}))$ and arrives at $(x_2(\tau_\ell), y_2(\tau_\ell))$;
- $\dot{x}_1(\ell) = \frac{x_1(\tau_\ell) - x_1(\tau_{\ell-1})}{\tau_\ell - \tau_{\ell-1}}$, $\dot{x}_2(\ell) = \frac{x_2(\tau_\ell) - x_2(\tau_{\ell-1})}{\tau_\ell - \tau_{\ell-1}}$, $\dot{y}_1(\ell) = \frac{y_1(\tau_\ell) - y_1(\tau_{\ell-1})}{\tau_\ell - \tau_{\ell-1}}$, and $\dot{y}_2(\ell) = \frac{y_2(\tau_\ell) - y_2(\tau_{\ell-1})}{\tau_\ell - \tau_{\ell-1}}$. The value $(\dot{x}_1(\ell), \dot{y}_1(\ell))$ and $(\dot{x}_2(\ell), \dot{y}_2(\ell))$ are, respectively, the pairs of the component-wise half widths of the segment that starts from $(x_1(\tau_{\ell-1}), y_1(\tau_{\ell-1}))$ and arrives at $(x_1(\tau_\ell), y_1(\tau_\ell))$, respectively, of the segment that starts from $(x_2(\tau_{\ell-1}), y_2(\tau_{\ell-1}))$ and arrives at $(x_2(\tau_\ell), y_2(\tau_\ell))$.

Distance in Eq. 2 can be naturally decomposed for sub-trajectories. For example, let's consider a normalized time-stamp equal to 0.25 and let's cut P_1 and P_2 trajectories, which share the same normalized time-stamps, at $\tau = 0.25$, we obtain two pair of sub-trajectories. We thus we can observe how the total squared distance is

¹ The trajectory is on a plane, but the extension to 3D spaces is straightforward.

decomposed in the squared distance of the first quarters (in terms of standardized time) of the trajectories and the last three quarters ones. We will consider this simple decomposition property in the following.

In a k-means like algorithm, it is important to define an average object. Indeed, k-means algorithms rely on the definition of a within cluster homogeneity criterion that usually is expressed as a distance between objects and a representative of the cluster. Once chosen a K number of clusters for partitioning N input objects, such that $\mathcal{P} = \{C_1, \dots, C_K\}$ is a partition of objects into K disjoint groups, and $G = \{g_1, \dots, g_K\}$ a set of average objects (or cluster prototypes), k-means like algorithms are based on the minimization of a criterion function $W(K, P)$ based on a distance $d(o_i, g_k)$ between a generic input object and the prototype of cluster C_k :

$$W(K, \mathcal{P}) = \sum_{k=1}^K \sum_{i=1}^N \mathbf{I}_{C_k}(i) \cdot d(o_i, g_k) \quad \text{where} \quad \mathbf{I}_{C_k}(i) = \begin{cases} 1 & \text{if } i \in C_k \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

We assume that given N trajectories on the 2D plane, registered such that they share a same normalized vector of increasing time-stamps τ having L elements, it is possible to define an average trajectory $\bar{P} = \{ \{(\bar{x}_0, \bar{y}_0), 0\}, \dots, \{(\bar{x}_L, \bar{y}_L), \tau_L\}, \dots, \{(\bar{x}_L, \bar{y}_L), 1\} \}$ where:

- $\bar{x}_L = \frac{1}{N} \sum_{i=1}^N x_L^i$ and $\bar{y}_L = \frac{1}{N} \sum_{i=1}^N y_L^i$

Having defined a squared distance and the average object consistently with the distance, it is possible to define a criterion W_B for a base k-means of trajectories as follows:

$$W_B(K, \mathcal{P}) = \sum_{k=1}^K \sum_{i=1}^N \mathbf{I}_{C_k}(i) \cdot d_E^2(P_i, \bar{P}_k) \quad \text{where} \quad \mathbf{I}_{C_k}(i) = \begin{cases} 1 & \text{if } i \in C_k \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Clustering trajectories using the base algorithm may leave hidden some interesting patterns in data. For example, if one imagines people waling in store, they have some common paths when entering or exiting from the store surely, due to access and exit doors, but different behaviors when they are inside the store. So, it could be interesting to consider different phases of an object trajectory, for example, an initial, a central and a final phase that could be delimited by two normalized time-stamps. In general, if we consider M sub-trajectories given from a common set of cuts depending on the choice of M normalized time-stamps, distance in Eq. 2 can be rewritten as follows:

$$\begin{aligned} d_E^2(P_1, P_2) &= \sum_{m=1}^M \int_{v_{m-1}}^{v_m} \left[|x_1(\tau) - x_2(\tau)|^2 + |y_1(\tau) - y_2(\tau)|^2 \right] d\tau = \\ &= \sum_{m=1}^M (v_m - v_{m-1}) \cdot \int_0^1 \left[|x_1(f_m(\gamma)) - x_2(f_m(\gamma))|^2 + |y_1(f_m(\gamma)) - y_2(f_m(\gamma))|^2 \right] d\gamma = \\ &= \sum_{m=1}^M (v_m - v_{m-1}) \cdot d_E^2(SP_{1m}, SP_{2m}) \\ &\text{where } f_m(\gamma) = v_{m-1} + \gamma \cdot (v_m - v_{m-1}) \end{aligned} \quad (5)$$

Using the distance in Eq.5 into Eq. 4, obviously it does not change the minimization problem and the results, giving a common initialization of centers, will be the same. A first proposal is to consider the trajectories as a set of M sub-trajectories, but without the term $(v_{m-1} - v_m)$. In this case a first modified k-means criterion W_S is the following:

$$W_S(K, \mathcal{P}, v) = \sum_{k=1}^K \sum_{i=1}^N \sum_{m=1}^M \mathbf{I}_{C_k}(i) \cdot d_E^2(SP_{im}, \overline{SP}_{km}) \quad (6)$$

The criterion in Eq. 6 is like to do a k-means on a multivariate dataset. In that case, each variable plays the same role in the clustering process. In terms of trajectory clustering, each set of sub-trajectories is equally important in the clustering process even if they have different variability (namely, sum of intra cluster squared distances w.r.t. the cluster prototype). Similarly to classical clustering, this effect can mask some cluster structures in the data. Starting from the approach of clustering using adaptive distances [3], that extends a k-means like algorithm (also known as dynamic clustering algorithm [2]), we propose to use cluster-wise adaptive distances for the k-means of trajectories.

Adaptive distances are weighted distances applied to cluster analysis. The use of adaptive distances in clustering allows automatically computing of weights that can be considered as relevance weights of each variable in a standard dataset. Two main cases are considered in [3]:

- Global weighting, namely, a weight for each variable is considered for all the data. In our case we consider a single weight for each set of sub-trajectories. The W_{AG} criterion function depends on M weights denoted with λ_m and is as follows:

$$W_{AG}(K, \mathcal{P}, v, \Lambda) = \sum_{k=1}^K \sum_{i=1}^N \sum_{m=1}^M \mathbf{I}_{C_k}(i) \cdot \lambda_m \cdot d_E^2(SP_{im}, \overline{SP}_{km}) \quad (7)$$

The minimization of W_{AG} has a trivial solution when all λ_m 's are null. For avoiding this trivial solution, in [3] the criterion function is minimized under the following constraint: $\lambda_m > 0$ and $\prod_{m=1}^M \lambda_m = 1$.

- Cluster-wise weighting, namely, a weight for each variable and each cluster is considered for all the data. We consider a weight for each set of sub-trajectories and each cluster. The W_{AL} criterion function depends on $K \times M$ weights denoted with λ_{km} and is as follows:

$$W_{AL}(K, \mathcal{P}, v, \Lambda) = \sum_{k=1}^K \sum_{i=1}^N \sum_{m=1}^M \mathbf{I}_{C_k}(i) \cdot \lambda_{km} \cdot d_E^2(SP_{im}, \overline{SP}_{km}) \quad (8)$$

For avoiding this trivial solution, in [3] the criterion function is minimized under the following constraint: $\lambda_{km} > 0$ and K constraints $\prod_{m=1}^M \lambda_{km} = 1$ for $k = 1, \dots, K$.

In the following section, we show the four algorithms variants according to the proposed criteria.

3 K-means clustering algorithms

In this section, we show the main steps of the four algorithms: trajectory k-means (TKM), sub-trajectory k-means (STKM), sub-trajectory adaptive k-means global (SKADAG), and cluster-wise (SKADAL). The TKM and STKM is the classic implementation of the k-means algorithm using criterion in Eq. 4, respectively in Eq. 6. The algorithms are initialized several times and the solution with a lower final criterion value identifies the best one. SKADAG and SKADAL algorithms are k-means algorithm with a new step that computes relevance weights automatically. The steps of the algorithms are the following:

- 0. Input:** A dataset P of normalized and registered trajecories cut at some pre-defined M normalized time-stamps. A predefined K number of clusters.
- 1. Initialization** Set $t = 0$
 - 1.1 Centers selection** Select randomly K trajectories and store them in $G^{(0)}$
 - 1.2 Fix initial weights** Fix $\Lambda^{(0)} = \mathbf{1}$
 - 1.3 Assign** Assign data to clusters according a minimum distance criterion, and generate the initial partition of trajetories $\mathcal{P}^{(0)}$
 - 1.4 Compute initial criterion** Compute $W_{AG}^{(0)}$ (SKADAG) or $W_{AL}^{(0)}$ (SKADAL)
- 2. Repeat** Set $t = t + 1$
 - 2.1 Centers selection** Fixed $\mathcal{P}^{(t-1)}$ and $\Lambda^{(t-1)}$, compute the average trajectories for each cluster and store them in $G^{(t)}$
 - 2.2 Compute weights** Fixed $\mathcal{P}^{(t-1)}$ and $G^{(t)}$, compute $\Lambda^{(t)}$ according to the constrained minimization of $W_{AG}^{(t)}$ (SKADAG) or $W_{AL}^{(t)}$ (SKADAL), using the Lagrange multiplier method.
 - 2.3 Assign** Fixed $G^{(t)}$ and $\Lambda^{(t)}$, assign trajectories to clusters according a minimum distance criterion w.r.t. the average trajectories, and store the partition of trajetories in $\mathcal{P}^{(t)}$.
 - 2.4 Compute the new criterion** Compute $W_{AG}^{(t)}$ (SKADAG) or $W_{AL}^{(t)}$ (SKADAL)
 - 2.5 Verify the stopping rule** If $W_{AG}^{(t)} < W_{AG}^{(t-1)}$ (SKADAG) or $W_{AL}^{(t)} < W_{AL}^{(t-1)}$ (SKADAL) then go to **2.** else go to **3.**
- 3. Return solution** Return $(P)^{(t)}$, $G^{(t)}$, $\Lambda^{(t)}$.

For the sake of brevity, we do not report the proof of the convergence to a local minimum of the algorithm, but it can be derived using the Lagrange multipliers method applied to the constrained minimizations of W_{AG} (SKADAG) or W_{AL} (SKADAL) for computing sequentially the best cluster averages, the best weights, and the best

partition once all the parameters, except the one of interest, are fixed. See [3], for further details.

4 Application

In this section, we present an application on two benchmark datasets of trajectory data the CROSS and the LABOMNI data, which have been used in ?? and are publicly available ². The CROSS dataset collects 1,900 cars trajectories approaching the crossroad in 19 (K) different ways. The LABOMNI dataset describes 15 (K) sets of trajectories of 209 people in a laboratory. In tab. 1, we report the accuracy of the proposed methods in recognizing the 19 patterns for the CROSS dataset and the 15 patterns for the LABOMNI one, using the Adjusted Rand Index (ARI), the purity index of clusters (PUR), and the Normalized Mutual Information Index (NMI), which are the most common external validity indexes used in classification problems.

Table 1 CROSS and LABOMNI datasets: trajectory clustering results. Cuts are reported for each dataset.

Methods	Datasets					
	CROSS			LABOMNI		
	N=1,900 k=19			N=209 K=15		
	ARI	PUR	NMI	ARI	PUR	NMI
K-means	0.8163	0.8389	0.9405	0.6715	0.8373	0.8248
	cuts (0.15,0.85)			cuts (0.005, 0.15 0.85,0.995)		
K-means pieces	0.8210	0.8411	0.9443	0.8772	0.9234	0.9118
K-means ADA global	0.8192	0.8400	0.9426	0.8930	0.9330	0.9230
K-means ADA local	0.8200	0.8405	0.9433	0.8273	0.9139	0.8998

From Tab. 1 we observe that algorithms based on sub-trajectories perform slightly better for the CROSS dataset and significantly better for the LABOMNI one. We remark that the two datasets have a different types of trajectories where, according to [7], CROSS datasets has less complex trajectories than the LABOMNI one. The complexity is related to the shapes of trajectories where those in the CROSS dataset are more regular, since they shows cars behaviors at a crossroad, while the LABOMNI dataset consists in trajectories of peoples that walk almost freely in a laboratory.

² http://cvrr.ucsd.edu/LISA/Datasets/TrajectoryClustering/CVRR_dataset_trajectory_clustering.zip

5 Concluding remark

We introduce new k-means like algorithms for trajectory data assuming that trajectories are normalized according to their time-stamps. We have used adaptive distances for taking into account the different relevance of sub-trajectories, and we have shown the usefulness of the proposed method in the application. Issues to be solved are mainly related to the optimal choice for cutting trajectories and the possibility of considering normalized trajectories or further information about, for example, the spatial and time length of the trajectories.

References

1. Demšar, U., Buchin, K., Cagnacci, F., Safi, K., Speckmann, B., Weghe, N.V., Weiskopf, D., Weibel, R.: Analysis and visualisation of movement: an interdisciplinary review. *Movement ecology*. **3**:5, (2015)
2. Diday, E.: The dynamic clusters method in nonhierarchical clustering. *International Journal of Computer and Information Sciences* 2: 61 (1973) doi: 10.1007/BF00987153
3. Diday, E. and Govaert, G.: Classification Automatique avec Distances Adaptatives. *R.A.I.R.O. Informatique Computer Science*, 11 (4), 329-349 (1977)
4. Ferreira, N. , Klosowski, J. T., Scheidegger, C. E. and Silva, C. T.: Vector Field k-Means: Clustering Trajectories by Fitting Multiple Vector Fields. *Computer Graphics Forum*, 32: 201-210. (2013) doi: 10.1111/cgf.12107
5. Jiang Bian, Dayong Tian, Yuanyan Tang, Dacheng Tao. A review of moving object trajectory clustering algorithms. *Artif Intell Rev* (2016) doi: 10.1007/s10462-016-9477-7
6. Lee, J., Han, J., Whang, K.: Trajectory clustering: a partition-and-group framework. *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 593-604 (2007)
7. Morris, B. T., Trivedi, M. M.: Learning Trajectory Patterns by Clustering: Experimental Studies and Comparative Evaluation, in *Proc. IEEE Inter. Conf. on Computer Vision and Pattern Recog.*, Maimi, Florida, (2009)
8. Sangalli, L.M., Secchi, P., Vantini, S., Vitelli, V.: K-mean alignment for curve clustering, *Computational Statistics & Data Analysis*, **54**,5, 1219–1233 (2010)

Bayesian Analysis of Privacy Attacks on GPS Trajectories

Analisi Bayesiana degli Attacchi alla Privacy su Traiettorie GPS

Sirio Legramanti

Abstract The success of applications for sharing GPS trajectories raises serious privacy concerns, in particular about users' home addresses. In this paper we show that a Bayesian approach is natural and effective for a rigorous analysis of home-identification attacks and their countermeasures, in terms of privacy. We focus on a family of countermeasures named "privacy-region strategies", consisting in publishing each trajectory from the first exit to the last entrance from/into a privacy region. Their performance is studied through simulations on Brownian motions.

Abstract *Il successo delle applicazioni per condividere traiettorie GPS solleva seri problemi di privacy, in particolare riguardo agli indirizzi di casa degli utenti. In questo articolo mostriamo come un approccio bayesiano sia naturale ed efficace per un'analisi rigorosa, in termini di privacy, dei tentativi di localizzare le abitazioni degli utenti e delle possibili contromisure. Ci concentriamo su una famiglia di contromisure chiamate "privacy-region strategies", che consistono nel pubblicare ciascuna traiettoria dalla prima uscita all'ultimo ingresso da/in una regione di privacy. Ne studiamo l'efficacia attraverso simulazioni su moti browniani.*

Key words: Harmonic measure, Home-identification, Privacy protection

1 Introduction

Nowadays GPS trajectories are not only easily recorded but also massively shared, for example via sport social networks. This raises serious privacy concerns, because from complete GPS trajectories it is possible to infer sensitive information about users, such as their home addresses [9, 4]. The problem can be tackled from two sides: on one hand, users should carefully select who can access their trajectories; on the other, the amount of sensitive information contained in GPS trajectories

Sirio Legramanti

Bocconi University, Milan, Italy, e-mail: sirio.legramanti@phd.unibocconi.it

should be limited. Since users' privacy awareness has been shown to be unreliable [8], the second approach is necessary, and can be implemented via algorithms taking GPS trajectories as input and giving as output modified trajectories from which it is harder to infer sensitive information. We call these algorithms *obfuscation strategies*, but they are also known as *location-privacy protection mechanisms* [12] or *trajectory privacy preservation mechanisms* [13]. Obfuscation strategies range from publishing nothing to publishing original trajectories as they are. Both extremes are undesirable in opposite ways that highlight the tradeoff between privacy and *utility*, which is the residual value of the obfuscated data for the considered application [13]. Publishing nothing guarantees perfect privacy but null utility, while full disclosure provides maximum utility but usually insufficient privacy.

In the present work, we focus on obfuscation strategies against *home-identification attacks*, in which an adversary tries to localize the house of a user exploiting his GPS trajectories. Krumm [7] and Hoh et al. [4] deal with the same type of attack but rely on heuristics giving as output, respectively, a single address and a list of addresses with no ordering in probability, thus failing to quantify the adversary's uncertainty, which instead we regard as a key part of a privacy measure. To overcome this issue, we propose a Bayesian framework for assessing the efficacy of privacy attacks and their countermeasures. Even if implemented for home-identification attacks on GPS trajectories, our framework can be used for other types of data and privacy attacks. A Bayesian approach is not completely new in privacy literature, but still mostly used to formulate privacy definitions [10, 1, 6]. A partial exception can be found in Shokri et al. [12], who however employ posterior bias as a privacy measure, completely disregarding the uncertainty quantification that naturally comes with a Bayesian approach.

2 Framework for Home-Identification Attacks

We model home-identification attacks on GPS trajectories as Bayesian inference problems in which the parameter of interest is the user's house location $\theta \in \Theta \subseteq \mathbb{R}^2$ and data consist of n published GPS trajectories $y^{(i)} = \{y_t^{(i)}\}_{t \in [0, \tilde{T}^{(i)}]}$ ($i = 1, \dots, n$), possibly altered before publication. We assume that the adversary has no access to the corresponding original trajectories $x^{(i)} = \{x_t^{(i)}\}_{t \in [0, T^{(i)}]}$ ($i = 1, \dots, n$), but knows the adopted obfuscation strategy and has some background knowledge about the user's house location θ , modeled by a prior distribution π . Under obfuscation strategy s and model m for the original trajectories, the published trajectories are then described by the hierarchical model

$$(y^{(i)} | x^{(i)}, \theta) \sim p_s, \quad (x^{(i)} | \theta) \sim p_m, \quad \theta \sim \pi \quad (i = 1, \dots, n).$$

We model both original and published GPS trajectories with continuous time stochastic processes even if GPS sensors actually record the position at discrete times, because the sampling frequency is usually high (a typical value is 1 Hz).

2.1 Privacy and Utility Measures

Since we model privacy attacks as point estimation problems, it is natural to measure privacy through the quality of the adversary's estimate: the better the estimate, the poorer the privacy. This principle holds for any privacy attack, while the type of sensitive attribute determines the loss function used to evaluate the estimate. Since in our case the sensitive attribute is the user's house location $\theta \in \Theta \subseteq \mathbb{R}^2$, we adopt a quadratic loss function. We then employ the posterior Mean Square Error (MSE) as a measure of the quality of the adversary's estimate, and hence as a measure of privacy. Through its well-known bias-variance decomposition, MSE incorporates both *correctness* and *uncertainty*, which are discussed as privacy measures in [12].

While the privacy measure depends on the sensitive attribute, the utility measure depends on the considered application. In the present work we focus on sharing GPS trajectories of fitness activities via sport social networks. In this context, fake times or locations are not allowed, since they would alter rankings. The only admitted form of perturbation is cutting part of the original trajectories. In particular, users may accept cuts to the initial and final part of their trajectories, which usually consist of how they get out of their neighborhood and back in. However, we assume that users want these cuts to be as small as possible. Based on these assumptions, the utility of a published trajectory $y = \{y_t\}_{t \in [0, \tilde{T}]}$ with respect to the corresponding original trajectory $x = \{x_t\}_{t \in [0, T]}$ can be defined as a monotone-decreasing function ϕ of the following *squared perturbation* (SP):

$$SP(y, x) = \begin{cases} \|y_0 - x_0\|^2 + \|y_{\tilde{T}} - x_T\|^2 & \text{if } y = x|_{[t_1, t_2]}, \text{ with } 0 \leq t_1 \leq t_2 \leq T \\ \infty & \text{otherwise} \end{cases} \quad (1)$$

where $x|_{[t_1, t_2]}$ denotes the restriction of x to $[t_1, t_2]$ and $\|\cdot\|$ the Euclidean norm in \mathbb{R}^2 . If the published trajectory is obtained by cutting the initial and/or final part of the original one, the SP is the sum of the square Euclidean distances between their starting and ending points, otherwise it is set to infinity, leading to the lowest possible utility.

The variability of original trajectories and the stochasticity of obfuscation strategies make the SP random. We then define the utility of obfuscation strategy s , under model m for original trajectories and prior π on θ , as

$$U_{m, \pi}(s) = \int \phi(SP(y, x)) dp_s(y | x, \theta) dp_m(x | \theta) d\pi(\theta). \quad (2)$$

In the present work, we do not need to quantify utility but just to set the same level of utility to fairly compare two strategies, hence we do not need to specify ϕ . A sufficient condition for two strategies to have the same utility, for any choice of ϕ , is to induce the same SP distribution.

2.2 Privacy-Region Strategies

As stated in §2.1, the only obfuscation strategies that are admissible for fitness GPS trajectories are those cutting at most the initial and final part of original trajectories. Since most of fitness activities either start and finish close to the user's house location θ or do not come close to it at all, this can be achieved by a class of strategies, that we call *privacy-region strategies*, cutting the part of each original trajectory before the first exit and after the last entrance from/into a privacy region containing θ . Each trajectory may be cut using a different privacy region, hence we denote with D_i the privacy region for the i -th trajectory. A privacy-region strategy s is described by

$$(y^{(i)} | x^{(i)}, D_i) = x^{(i)}_{[t_1^{(i)}, t_2^{(i)}]}, \quad (D_i | x^{(i)}, \theta) \sim p_s \quad (i = 1, \dots, n)$$

with $t_1^{(i)} = \inf\{t \in [0, T^{(i)}] : x_t^{(i)} \notin D_i\}$ and $t_2^{(i)} = \sup\{t \in [0, T^{(i)}] : x_t^{(i)} \notin D_i\}$. If the infimum for $t_1^{(i)}$ is over an empty set (i.e. the original trajectory never leaves the privacy region) then no trajectory is published and the SP is set to infinity.

The distribution of each privacy region D_i may depend on the corresponding original trajectory, allowing to design adaptive obfuscation strategies. In this work, though, we do not implement this dependency, leaving it to future research.

We now describe two privacy-region strategies, corresponding to different distributions of the privacy regions: *random-radius strategy* and *two-balls strategy*.

In a random-radius strategy of parameters $\alpha > 0$ and $\beta > 0$, the privacy regions are balls centered in the user's house location θ , with i.i.d. random radii r_i such that $r_i^2 \sim \text{Gamma}(\alpha, \beta)$ (see Figure 1a).

Remark 1. Other choices for the distribution of r_i are possible. The case $r_i \sim \delta_{r^*}$, with $r^* > 0$ fixed, leads to *fixed-radius strategy*. We will not analyze this strategy, even if used in industry, since an adversary knowing r^* can locate θ with no uncertainty based on just three distinct exit/entrance points from the privacy region, which in this case is almost surely unique, leveraging elementary geometry.

In a two-balls strategy of parameters $\alpha > 0$, $\beta > 0$, $r > 0$ and $R > r$, all the trajectories of the same user are cut employing as a privacy region the same ball $B(c, R)$, where the radius R is fixed while the center c is randomly distributed within $B(\theta, r)$ as follows: $c = \theta + r\rho[\cos \tau, \sin \tau]^T$, $\tau \sim U(0, 2\pi)$, $\rho^2 \sim \text{Beta}(\alpha, \beta)$ (see Figure 1b). This strategy generalizes the one proposed in [7], where c was uniformly distributed within $B(\theta, r)$, corresponding to $\alpha = \beta = 1$. The restriction $r < R$ ensures that θ lies within the privacy ball. A value of r too close to zero, though, would almost reduce the two-balls strategy to the fixed-radius strategy, that we have shown to be inefficient in Remark 1. Also ρ^2 concentrated near zero would produce the same effect. On the other extreme, ρ^2 concentrated near one corresponds to c concentrated near $\partial B(\theta, r)$. In this case, if moreover r is close to R , we have that θ will be close to the boundary of the privacy ball with high probability, producing exit points concentrated close to θ . Hence, small values of r and distributions of ρ^2 concentrated near zero or one should be avoided.

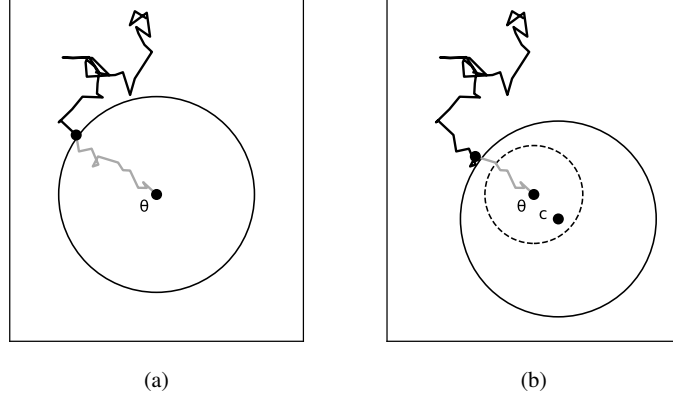


Fig. 1 Random-radius strategy (a) and two-balls strategy (b) on the same original trajectory. The unpublished part of the trajectory is plotted in grey, while the published part is in black. The privacy balls have solid border: for the random-radius strategy the privacy ball is centered in θ , while for the two-balls strategy it is centered in a random point c within the ball with dashed border.

2.3 Privacy-Region Strategies on Brownian Motions

Our framework poses no limitation to the complexity of the model for the original trajectories, but we start with one of the simplest possible models and assume that, given the user's house location θ , the original trajectories are independent Brownian motions starting at θ . Even if this is not a realistic model for human movement, it can be useful to rule out candidate obfuscation strategies. In fact, if a strategy is unable to hide the start of a Brownian motion, it cannot aim to hide the start of the much more structured human movement.

Thanks to the memoryless property of the Brownian motion, we can consider a simplified version of privacy-region strategies, consisting in cutting just the part of each original trajectory before the first exit from the corresponding privacy region:

$$(y^{(i)} | x^{(i)}, D_i) = x^{(i)}_{|_{[t^{(i)}, +\infty)}}, \quad (D_i | \theta) \sim p_s \quad (i = 1, \dots, n),$$

with $t^{(i)} = \inf\{t \in \mathbb{R}^+ : x_t^{(i)} \notin D_i\}$. The squared perturbation (1) also simplifies to $\|y_0 - x_0\|^2$. The exit time $t^{(i)}$ is a stopping time for $\{x_t^{(i)}\}_{t \in \mathbb{R}^+}$ [11, Remark 2.14] and, if the privacy region D_i is almost surely a bounded domain containing θ , then $t^{(i)}$ is almost surely finite [3, Chapter 4.3]. Moreover, conditional on the exit point of the original trajectory from D_i , the published part of the trajectory is a Brownian motion independent of the unpublished part [11, Theorem 2.16]. It follows that the exit points of the original trajectories from the corresponding privacy regions are sufficient statistics for θ . Given D_i and θ , each exit point z_i ($i = 1, \dots, n$) is dis-

tributed according to the harmonic measure of parameter θ on the boundary of D_i [5], a distribution that we denote with $\mathcal{H}_\theta^{D_i}$.

In conclusion, if the original trajectories are Brownian motions starting at θ , home-identification attacks against privacy-region strategies can be represented as Bayesian inference problems with the following hierarchical structure:

$$(z_i | D_i, \theta) \sim \mathcal{H}_\theta^{D_i}, \quad (D_i | \theta) \sim p_s, \quad \theta \sim \pi \quad (i = 1, \dots, n).$$

3 Simulation Studies

We consider illustrative simulations on Brownian motions to understand which strategy among random-radius and two-balls guarantees higher privacy. The comparison is fair only if the two strategies have the same utility [13]. Under our utility definition (2), a sufficient condition for this is having the same SP distribution. We relax this condition and constrain just the first two SP moments to be the same. Such moments are explicitly available only for the random-radius strategy, for which $SP \sim \text{Gamma}(\alpha, \beta)$. Hence, we fix the two-balls parameters first, then set the random-radius ones so that the first two theoretical SP moments of the random-radius strategy match the first two sample SP moments of the two-balls strategy.

We consider different sets of two-balls parameters, listed in Table 1. For each setting, we generate 50 Brownian motion trajectories starting at θ , which is fixed at the origin since the considered strategies are translation invariant. We then process these trajectories through both strategies, producing 50 cut trajectories each, on which we perform Bayesian inference for θ , under a uniform improper prior, using *PyStan* [14]. From the posterior samples of θ we compute a Monte Carlo estimate of the posterior MSE, plotted in Figure 2 for all settings in Table 1. We can observe that the MSE, which is our measure of privacy, is higher with the two-balls strategy than with the random-radius strategy, in all settings. Especially with the two-balls strategy, higher SP (i.e. lower utility) corresponds to higher MSE (i.e. higher privacy), in the expected privacy-utility tradeoff.

In Figure 3 we plot the posterior MSE as sample size grows, for a single setting ($r = 1$, $R = 3$, $\alpha = \beta = 4$). The MSE goes to zero under both strategies, but is stably higher with the two-balls strategy than with the random-radius strategy. We can also observe that matching the first two SP moments produces SP densities that are similar in shape as well, making the comparison particularly fair.

Finally note that each home-identification attack on 50 trajectories took about one second on a regular laptop (Intel Core i7-3632QM CPU @ 2.20GHz x 8, 7.7 GB of RAM), with running time scaling linearly in the number of trajectories. The modest amount of time and computational resources needed for the attack confirms that risks for users' privacy are real.

TB parameters					MSE median	
r	R	α	β	mean SP	TB	RR
1	3	4	4	8.34	0.25	0.03
1	4	4	4	15.34	0.48	0.04
2	5	4	4	23.23	0.69	0.13
1	5	4	2	24.29	0.69	0.05
1	5	4	4	24.42	0.78	0.04
1	5	2	4	24.71	0.77	0.02

Table 1 Settings considered for the comparison between the two-balls strategy (TB) and the random-radius strategy (RR).

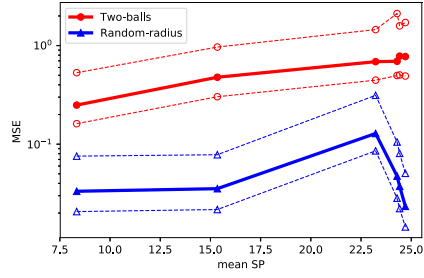


Fig. 2 Posterior MSE given 50 trajectories for the settings in Table 1. The median is plotted solid, while 5% and 95% quantiles are dashed.

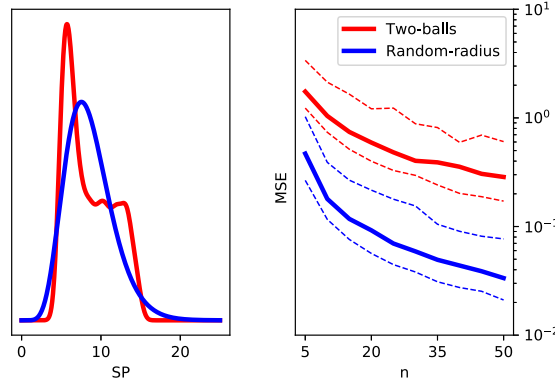


Fig. 3 Two-balls strategy vs. random-radius strategy as sample size grows, for a single set of two-balls parameters ($r = 1$, $R = 3$, $\alpha = \beta = 4$). On the left, the SP pdf; on the right, the posterior MSE median (solid) and its 5% and 95% quantiles (dashed) as functions of the sample size n .

4 Discussion

Further analysis of privacy-region strategies may involve the development of more realistic models for GPS trajectories, which can still be plugged in our framework. This may lead to the loss of some of the properties exploited here, as the sufficiency of exit points or the availability of their likelihood. Anyway, as long as a generative model and some quasi-sufficient statistics are available, Bayesian inference can still be carried out through Approximate Bayesian Computation [2]. Exit points, which are sufficient statistics in the Brownian motion case, are the first candidates as quasi-sufficient statistics under more general models.

Other possible extensions of the present work are represented by different applications, obfuscation strategies or even data-types. In fact, the utility measure is the

only application-specific element in our framework, which in all the other aspects is completely general.

Acknowledgements I would like to thank Giacomo Aletti, Daniele Durante, Nelly Litvak and Lucia Paci for their helpful comments on a first version of this work.

References

- [1] Bassily, R., Groce, A., Katz, J., and Smith, A. (2013). Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *2013 IEEE 54th Annual Symp. on Found. of Computer Science*, pages 439–448.
- [2] Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.
- [3] Chung, K. L. (2013). *Lectures from Markov processes to Brownian motion*, volume 249. Springer.
- [4] Hoh, B., Gruteser, M., Xiong, H., and Alrabady, A. (2006). Enhancing security and privacy in traffic-monitoring systems. *Pervasive Computing*, 5(4):38–46.
- [5] Kakutani, S. (1944). Two-dimensional Brownian motion and harmonic functions. *Proc. of the Imp. Acad.*, 20(10):706–714.
- [6] Kasiviswanathan, S. P. and Smith, A. (2014). On the ‘semantics’ of differential privacy: A Bayesian formulation. *J. of Privacy and Confidentiality*, 6(1):1–16.
- [7] Krumm, J. (2007). Inference attacks on location tracks. In *Int. Conf. on Pervasive Computing*, pages 127–143. Springer.
- [8] Krumm, J. (2009). A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399.
- [9] Liao, L., Fox, D., and Kautz, H. (2006). Location-based activity recognition. In *Adv. in Neural Information Processing Systems*, pages 787–794.
- [10] Machanavajjhala, A., Gehrke, J., and Götz, M. (2009). Data publishing against realistic adversaries. *Proc. of the VLDB Endowment*, 2(1):790–801.
- [11] Mörters, P. and Peres, Y. (2010). *Brownian motion*. Cambridge Univ. Press.
- [12] Shokri, R., Theodorakopoulos, G., Le Boudec, J.-Y., and Hubaux, J.-P. (2011). Quantifying location privacy. In *2011 IEEE Symp. on Security and Privacy*, pages 247–262.
- [13] Singh, R., Theodorakopoulos, G., Marina, M. K., and Arapinis, M. (2018). On choosing between privacy preservation mechanisms for mobile trajectory data sharing. In *2018 IEEE Conf. on Commun. and Network Security*, pages 1–9.
- [14] Stan Development Team (2017). Pystan: the Python interface to Stan, Version 2.16.0.0. <http://mc-stan.org>. Online: accessed March 12, 2019.

Data Analytics in the Insurance Industry:

Market trends and lessons from a use case customer predictive modelling

Data Analytics nel settore assicurativo: principali trend e considerazioni da un caso d'uso applicato alla predizione del comportamento degli assicurati

Cristian Losito and Francesco Pantisano

Abstract This paper discusses the impact of Big Data and advanced analytics concepts in the Insurance Industry, both in terms of emerging market trends and some practical applications. A recent EY use case is presented, in which linear and non-linear models are applied to predict customer behaviour on a pension product.

***Abstract** Il presente documento affronta il tema Big Data e l'utilizzo di soluzioni di advanced analytics nel settore Assicurativo, sia presentando i principali trend di mercato che le applicazioni. E' illustrata una recente esperienza EY, in cui modelli lineari e non lineari sono stati applicati per predire il comportamento degli assicurati di un prodotto pensionistico.*

Key words: Insurance, Big Data, Advanced Data Analytics, Predictive Modelling.

1 Introduction

The insurance industry is facing major disruptive changes to existing business models due to digital technologies and the data deluge that these bring.

These changes represent both opportunities and threats to current market players, while we are also observing the rise of new Insurtech players which occupy various positions on the value chain of the industry. Furthermore, sectoral convergence is driving the creation of

¹

Cristian Losito, IOA & FIA; cristian.losito@it.ey.com

Francesco Pantisano, PhD; francesco.pantisano@it.ey.com

digital ecosystems within which insurance covers are offered as a “service” rather than a stand-alone product.

This paper discusses the impact on the Insurance Industry of some macro trends brought on by digitalisation and big data, namely:

- Customer interactions & Distribution models,
- The use of big data in pricing,
- Predictive modelling and Fraud detection.

In the following sections, each of the above is explored in turn. Finally, a practical “real-life” application of data analytics in the insurance industry is presented.

2 Digital Disruption and market trends in the Insurance Industry

2.1 Customer interactions & Distribution models

One of the main macro-trends connected to the rise of digital technologies is the greater importance of “digital savvy” customers in the market place. These customers often have different “insurance needs” and higher “experience” expectations compared to the past. This is driving changes to the way existing insurance companies are looking to interact with their customers as well as creating a market for new entrants.

Examples of such changes can be seen in the way in which insurance products are distributed and how insurers interact with customers. Customers are demanding faster interactions via intuitive and customer friendly interfaces.

This has seen the rise of online services and mobile apps that manage the entire cycle of the customer journey from quotes, underwriting, insurance cover check-ups, policy statements and claims management. Traditional players in the market are increasingly providing such services, but we have also witnessed new entrants with agile business models that often focus on “gaps” in the value chain. Firms such as Lemonade [1] have emerged in this very space.

Evolving perceptions of insurance needs, coupled with advances in digital technologies, have also led to instant insurance models which aim to provide specific covers (e.g. for skiing accidents) for a limited timeframe (e.g. daily). Currently most of these contracts have to be actively purchased by the customer via web or mobile applications, but the next frontier is for “behavioural models” to identify pre-emptively the insurance need (e.g. geolocation on mobile devices can identify the proximity to a skiing resort and an insurance app can offer skiing cover via push notifications directly on the smartphone).

Another technological enabler for innovation in insurance is the use of “wearables” (e.g. digital wristbands or smart watches) that can be used to collect biometric parameters. This data is often used for recreational activities (e.g. tracking daily steps, run times and distances) and is then displayed on personalised dashboard or connected to medical services. Wearable devices can be used to monitor the health of an individual in order to act as warning system for certain conditions thus favouring early intervention rather than a costly insurance claim. For example, Chronolife [2] has developed a “smart-vest” for individuals prone to chronic or congestive heart failure which measures six physiological stats via sensors. These sensors collect data in real time and use machine learning to predict the likelihood of an oncoming heart attack. Similarly, big data can be collected from Internet of Things (IoT) devices (e.g., smart-meters or smart-home devices) which can be used to better identify and manage insurance needs (e.g. home insurance cover).

2.2 The use of big data in pricing

The onset of digital technologies have also brought about another great driver for change: Big Data.

Traditionally, insurance companies and mutual societies have always served society by providing a vehicle to collectively share large losses resulting from an accident or life-changing event which for a single individual or business would otherwise prove unaffordable. This is done by pooling together the financial resources of the “insured community” and the relative contribution of each individual or business to this pool takes the form of affordable insurance premiums. The role of the pricing actuary in this context is to determine the insurance premium allowing for the underlying risk associated with the insurance cover, the expenses and the profitability margins required by the enterprise.

Ever since the dawn of the actuarial profession, the core activity of calculating the “correct” insurance premium has fundamentally relied on empirical statistics of the insurable event (e.g. mortality tables, catastrophe events for crops, etc.) and associated factors that influence the level of risk. It should come to no surprise that the greater access to data available from a variety of sources, ranging from wearables, apps, or social media has provided a platform to enhance pricing technicalities.

Some notable examples of big data sources for pricing purposes include:

- *Genetic-testing data*: One emerging development in life and health pricing is the use of genetic data to reduce the risk of adverse selection. Genetic testing is becoming increasingly available at lower prices and their use for life, disability, critical illness and long-term care insurance is expected to increase in relevance, in the next few years, as testing becomes widespread and more detailed clinical data emerges [3].
- *Car telematics*: the development of sensors installed on cars are used to collect data which enables the driving habits and behaviour of the driver to be analysed. This type of dynamic pricing model is referred to as *pay-how-you-drive* [4] and enables factors which express the driving style (e.g. acceleration & breaking, speed, etc.) and route patterns (e.g. heavy traffic routes, driving at night, driving conditions, street parking in high crime areas, etc.) to be allowed for when pricing the insurance cover.

The examples above show how big data can be used to better profile the risk of the individual and thus determine the level of cover that is tailored to the consumer’s needs and a corresponding insurance premium that allows for individual pricing factors. However, while increases in the availability of big data can help to better value risks, there are some ethical considerations regarding the social cost that this may bring.

Indeed, General Data Protection Regulation (GDPR) has recently come into force in a bid to safeguard the treatment of personal data. Nevertheless, there continues to be a lot of scrutiny regarding data privacy and the role of social media companies in this regards. How this issue evolves will have repercussions on the extent that this data can be used by insurers.

There is also an argument that personalised pricing would ultimately lead to a breakdown of the mutuality principle upon which insurance, from a social perspective, is founded upon. This would essentially lead to bad risks being offered unaffordable insurance terms and thus de-facto excluding those individuals from accessing insurance cover. For insurances services that are deemed integral to society this may be undesirable in the eyes of governments and regulators alike who aim to safeguard the general public interest.

This is highly relevant with regards to genetic data. Although such information offers undoubted advantages from a risk pricing perspective (it is unique for each person, it has a

high predictive power, etc...) it may also be deemed discriminatory or lead to emotional damages for those individuals classed as high risk who may find it difficult to obtain cover. Some medical practitioners affirm that the Insurer right to ask for a genetic test could lead people to postpone or even cancel the test (in the fear of losing the opportunity to buy a policy) impacting negatively on their wellbeing. On this aspect, existing regulations generally apply to employment and health insurance, while for life, disability, critical illness and long-term care insurance, regulations and social attitude are quite diversified across territories.

2.3 Predictive modelling and Fraud Detection

Big Data techniques can also be used for internal processes within Insurance companies in order to render more efficient the use of internal data assets and improve business steering decisions.

Examples of such applications include:

- *Predictive modelling*: Machine learning can be applied to predict policyholders' behavioural patterns in relation to decisions at their discretion. This may include the definition of policy lapses or renewal, as well as increase or decrease in the policy coverage, or the definition of upselling options. A firm that better understands the behavioural patterns of its own customers can leverage this information to improve customer satisfaction and company profitability. Practical applications include:
 - providing a profile of a “desirable customer” to the sales network in order to render targeted marketing campaigns more effective (for example as part of an up-sale or cross-sale initiative)
 - providing an identikit of policyholder likely to lapse a contract to retention teams
 - nudge policyholders' behaviour towards mutually-desirable decisions (e.g. elect to continue paying into a pension plan)
- *Fraud detection*: the cost of undetected frauds is reflected in the insurance premium as higher expected claims, and thus, is ultimately borne by policyholders. Insurance companies traditionally assess claims through a series of controls ranging from the use of standardised checks (e.g. via a questionnaire) to detailed inspection by experienced claim handlers. However, standard approaches prove only partially effective, as they are mostly based on the claim handlers' experience and judgement, resulting in costly, time-exhausting claim processes. Instead, machine learning algorithms can *learn* from past fraud patterns and *evaluate* new claims by flagging-up cases of suspicious features and streamlining valid claim settlements. Furthermore, the data can be augmented with additional sources (e.g. weather data or social network data) or image recognition techniques (e.g. photo-based damage assessment) in order to significantly improve fraud detection capabilities of an insurance company. Ultimately a more efficient and effective claim handling process leads to better customer service and satisfaction.

3 Use Case: Machine learning for predicting pension premium patterns

In this section, we explore a real life use case of machine learning applied to predict customer behaviour. The use case in question relates to a defined contribution pension product portfolio [5] of a large European insurance company. The pension product is designed so that the

Data Analytics in the Insurance Industry

policyholder pays recurrent premiums (through a combination of employer and voluntary contributions) up to retirement age, after which they become eligible for a retirement benefit which may take the form of either a capital sum, an annuity or a combination of both. The pension contributions are tax deductible (up to a limit of €5,164 per annum) so there is a fiscal incentive to the policyholder to paying the premiums.

The company had experience higher than expected pay-up rates (i.e. non-renewal), and this therefore meant that they had lower premium income compared to their expectation. In order to investigate the underlying factors of this phenomenon a predictive model was deployed. The objective was to identify the customer profile that is more likely to adhere to the premium paying plan and those at risk of paying up. The output were used to improve retention strategies of the existing customer base and provide insights for the redesign of the product to mitigate the risk of high pay-up rates in the future.

3.1 Methodology

Firstly, it is important to contextualize the problem and recognize that the nature of the product provides the customer with a high level of flexibility in terms of deciding when and how much to pay into the pension plan. From a modelling perspective, this is very complex since we are trying to predict a pattern rather than a specific event (as would be the case for lapses for example).

In order to approach the problem, the first step was to define six states that represent the possible premium paying patterns over the course of a 12 month period:

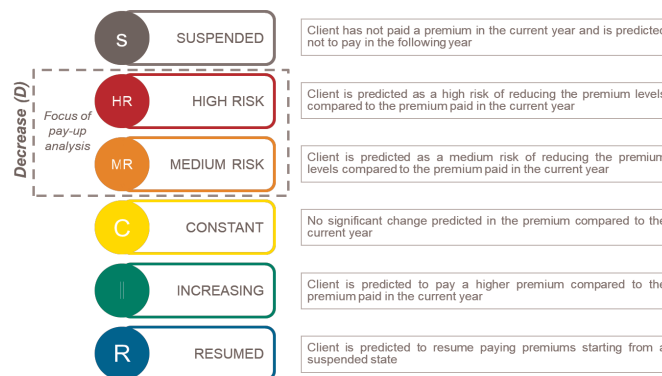


Figure 1: Premium paying states of the model

The second stage was to construct a data model of observed data. This was initially based on a set of 160+ sector-specific features of which 28 emerged as main predictors (see Sec 3.2). The main categories of observed features are:

- *Descriptive features:* include all the relevant demographics characteristics of the policyholder (e.g. age, profession, location) and sales channel information (e.g., agency size and revenue);
- *Volume-related features:* includes total assets under management for the current policy, assets under management for other savings contracts, current and past premium contributions, etc.;

- *Policyholders' behavioral features*: variables that represent observed trends (e.g., number of consecutive premiums paid, number of interruptions, etc.).

An important element of the data model was to utilize transaction data from other savings, protection and general insurance contracts that the policyholder had with the company, in order to build a holistic picture of the behavioural pattern of the customer.

The third step was to construct the predictive model. For this purpose a waterfall structure was selected that relied on a combination of a Random Forest technique to model the most extreme behaviour (i.e., Suspended, High-Risk and Resumed states), while a linear classification model was used for Medium Risk, Constant and Increasing states.

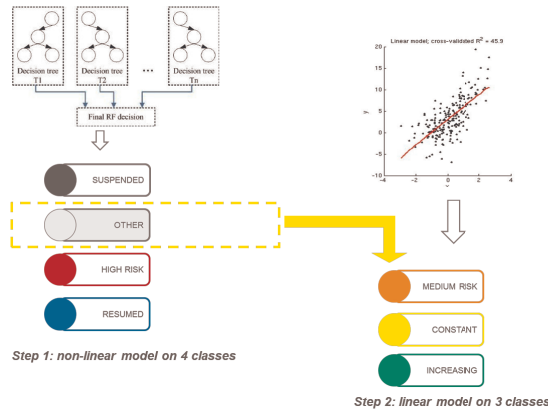


Figure 2: Two-step waterfall analytical model

The random forest model was calibrated with an ensemble method using random selection of variables and bootstrap samples. Accordingly, recursive partitioning trees were developed by using majority vote and each tree was based on a bootstrap sample of the training set. A number m is specified to be much smaller than the total number of variables M . Therefore, at each node, m variables are selected at random out of the M , and the split is the best split on these m variables [7-8]. The most powerful aspect about random forests is the variable importance ranking which estimates a predictive value by scrambling the variables and evaluating the performance drop with respect to the model performance. Moreover, when applied to the data, random forests outperformed traditional logistic regression-based scores due to the interaction effects across data that are predictive.

The model has been calibrated by splitting the data between training and validation set [4].

3.2 Results

In order to optimize performance of the predictive model, starting from the total set of features that were initially considered in the data model, we selected those that were highly descriptive of the policyholder's behaviour. The relevance of predictive features for the Random forest is measured by Gini value (Figure 3 below) while the relevance of predictive features of the linear model is indicated by their p -value. A predictor with a low p -value is likely to be a meaningful addition to your model, since changes in the predictor's value are related to changes in the response variable.

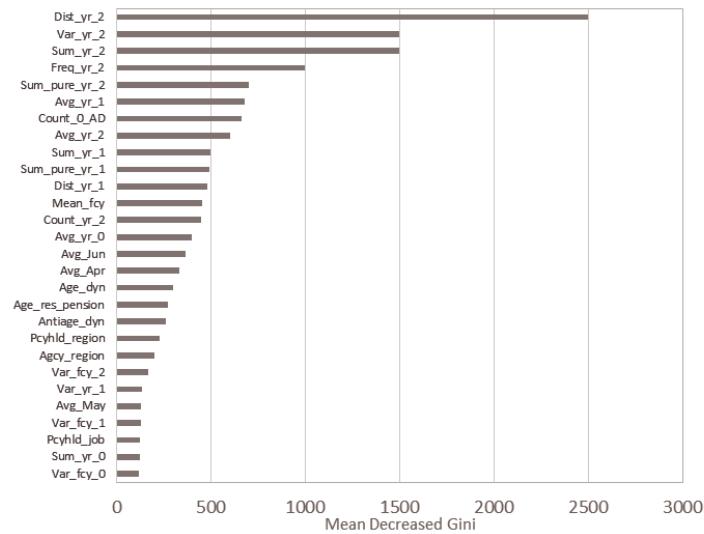


Figure 3: Gini coefficient for main predictors.

The accuracy of the model is assessed via a confusion matrix (Table 1 below), in which columns and rows respectively refer to actual observed states and model predictions.

Table 1: Confusion matrix for the Predictive Model

<i>Confusion Matrix</i>	H/M Risk	Resumed	Suspended	Constant	Increasing	Sum
High/Medium Risk	1857	0	0	1661	515	4033
Resumed	0	65	91	0	0	156
Suspended	0	924	9617	0	0	10541
Constant	1235	0	0	5271	664	7170
Increasing	195	0	0	989	177	1361
Sum	3287	989	9708	7921	1356	23261

The overall accuracy of the model is 73%, which outperforms a traditional clustering model by a factor of 2. The model performs particularly well in predicting the Constant and Suspended states with an accuracy comprised between 70-90%. The transition from Suspended to the Resumed state (i.e. those policies that resume paying after a suspension period) are harder to capture due to the scarce to no availability of indications for such phenomena among training data. Finally, the model accurately predicts around half of the high-to-medium risk class.

Furthermore, a backtesting analysis was performed by comparing the distribution of predicted states with the actual observed behaviour over the subsequent 12 months.

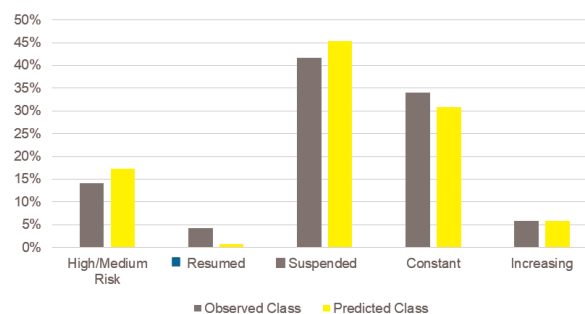


Figure 4: Backtesting prediction against actual data

Some of the main considerations regarding the predictive power of the model features are:

- Recent behaviour is the most important predictive feature. Those that experienced recent irregularities in their premium contributions were more likely to decrease or stop their premiums.
- Premium size: policyholder paying premiums of around €1,000 are more likely to stay in the constant state compared to lower and higher level of premiums.
- Premium frequency: policyholder paying monthly premiums are three times more likely to continue paying compared to those paying annual premiums.
- Direct debit policyholders are twice more likely to continue paying.
- Socio-demographic features such as profession, location and age have a moderate predicted power. For example, those aged 20-30 are more likely to interrupt their pension contributions due to other competing financial needs.
- Agency size: larger agencies are associated with more stable premium patterns.

4 Outlook & Conclusion

This paper presented opportunities and challenges related to Big Data and advanced analytics techniques in the insurance industry. These emerging trends are gaining momentum, revolutionising applications in customer intelligence and pricing models.

References

1. Lemonade Insurance Company – <http://www.lemonade.com>
2. Chronolife Real Time Insurance - <http://www.chronolife.net>
3. Code on Genetic testing and Insurance, ABI & HM Government 2018
4. Introducing 'Pay-How-You-Drive' (PHYD) Insurance, EY, 2016
5. Blake, D., Pension Economics. John Wiley & Sons, (2006)
6. Gerber, H. Life Insurance Mathematics, third ed. Springer, (1997)
7. Assmussen, S., and Glynn, P.: Stochastic Simulation: Algorithms and Analysis. In Springer, (2007).
8. McNeil, A., Frey, R., and Embrechts, P. Quantitative Risk Management, Princeton (2015)

BasketballAnalyzeR: the R package for basketball analytics

BasketballAnalyzeR: il pacchetto R per l'analisi dei dati nella pallacanestro

Marica Manisera, Marco Sandri and Paola Zuccolotto

Abstract The aim of this contribution is to introduce the upcoming R package `BasketballAnalyzeR`, which allows to perform basic and advanced performance analysis in basketball.

Abstract *L'obiettivo del contributo è di presentare le funzionalità del pacchetto R `BasketballAnalyzeR`, che sarà rilasciato a breve, per effettuare analisi della performance (base e avanzata) nella pallacanestro.*

Key words: sports analytics, performance analysis, basketball data science, NBA analytics

1 Introduction

The use of statistical methods and models in sports has rapidly grown in the last decades, as the huge number of books and articles published on scientific and specialized journals demonstrates [1, 2, 7, 11]. Results from sports analytics are more and more used as a support for the strategic decisions of teams and athletes, and numbers and quantitative information fill television broadcasts, web pages and posts on social media for the amusement of fans. A quantitative approach increasingly characterizes all sports in several fields: performance analysis, study of the psychological traits of athletes, marketing communication, sport management, finance of the sport societies, sentiment analysis of the fans, and many others.

Marica Manisera
University of Brescia, C.da S. Chiara, 50 - 25122 Brescia, Italy. e-mail: marica.manisera@unibs.it

Marco Sandri
DMS-StatLab, University of Brescia, C.da S. Chiara, 50 - 25122 Brescia, Italy.
e-mail: sandrimarco@gmail.com

Paola Zuccolotto
University of Brescia, C.da S. Chiara, 50 - 25122 Brescia, Italy. e-mail: paola.zuccolotto@unibs.it

In our contribution, the focus is on performance analysis in basketball, in an era in which most coaches and their backroom staff rely on formulas and figures to predict the most effective methods for winning. This is especially true in the US, where NBA is leading the major transformation related to the use of analytics: experts rely on data to measure a team's probability of winning and to assess a player's or a team's performance based on the "eye-test", that is the impression that came from watching a game, is out of fashion. The attention is in particular on basketball analytics used to analyze performance issues related to players and teams, playing patterns, game strategies, performance drivers, interactions among players.

The aim of this contribution is to describe some features of the forthcoming R package `BasketballAnalyzeR` [9], specifically designed to accompany the upcoming book entitled "Basketball Data science", proposed for addressing the most recent advances on basketball analytics, with an essentially empirical approach. The book has been developed within the international project BDsports (Big Data analytics in sports, bdsports.unibs.it), whose main aims include scientific research, education, dissemination and practical implementation of sports analytics.

After a brief recall to the state of the art in basketball analytics, Section 2 introduces different data sources that can then be used to perform both basic (Section 3) and advanced analyses (Section 4).

The main functionalities of `BasketballAnalyzeR` will be presented with the help of real data examples. Starting from concrete research questions, posed by basketball experts, we will show how to use the package `BasketballAnalyzeR` in order to find an answer based on selected statistical methods, data mining algorithms and visualization tools.

2 State of the art of basketball analytics and basketball data

To simplify, basketball analytics can be viewed as composed of three main worlds: institutional analyses, sport analytics services and scientific research. Analysts hired by teams should be considered as a special case, because their data and analyses are usually top secret.

Institutional analyses involves all the teams' statistics available on the web, in real time, for free. Many basic statistics for teams and players are usually provided for every league and every championship: they include success percentage in 2-point shots, 3-point shots, free throws, number of shots, number of fouls or rebounds, turnover, and so on. This information is useful but not enough to make the difference in defining game strategies or predicting winning probabilities.

The second world of basketball analytics is related to sport analytics companies, which usually create platforms and services, most of times customized, for analyzing data and allowing teams to choose the best solutions and strategies. They usually offer support in real time. Among other activities, sport analytics companies develop applications easy to use that allow to register all the events of a match and provide

a lot of outputs, including visualization tools, to help the team's decisions. Sport analytics companies enter the business side of the sports industry.

Third, basketball analytics can be done by scientific research, which aims to perform more sophisticated analyses, using up-to-date statistical methods and models. Complex results are obtained, and they should be evaluated with the help of basketball experts, which prevents their use in real time. There exist scientific journals aimed at publishing research on the quantitative aspects of sports, in order to give a response to the increasing demand for novel methods to analyze and understand data in the growing field of sports analytics. In addition, statistical journals not mainly focused on sports publish articles or special issues on sports analytics [10, 3, 4]. Many papers about basketball analytics have been published, with a variety of aims, from a simple description of the game to more complex problems, such as forecasting the outcomes of a game, analyze players' performance, studying the network of players, and how each player relates to his teammates, their spatial positioning, and the identification of optimal strategies.

The range of possible questions that may be answered by statistical analyses is growing thanks to the online availability of large data sets and increasing computational power. Data can be obtained by multiple sources, such as the National and International Federations, sporting organizations, basketball societies and others. Basketball data can be freely available or on payment and the collecting procedures can be different: usually, when data are for free, high-level computer procedures are needed to obtain data (for example, web scraping codes). In [6], data are classified in the following macrocategories: (i) data recorded manually (for example, basic aggregate statistics in the box scores or play-by-play data, recording all the events occurred during a match); (ii) data detected by technological devices (for example, positions of the players on the court, recorded by sensors or other tracking systems); (iii) data from psychological scales administered to players in order to measure their subjective traits (for example, leadership, coping strategies, ...); (iv) other data, including, for example, those retrievable from the web (posts of the fans on social media, trends of online searches from Google Trends and other tools).

BasketballAnalyzeR is presented using real datasets, available in the package and including box scores, play-by-play data and supplementary information about the teams. All data are real and come from the matches of the regular season of the NBA championship 2017/2018 (box scores refer to 1230 matches played by all teams, while play-by-play data concern the 82 matches played by Golden State Warriors).

3 Basic analytics

BasketballAnalyzeR allows to perform basic analytics on any team or player or match or tournament, given that appropriate data are available (box scores or play-by-play data). Basic analytics are based on the concepts recalled in the first scientific paper on basketball analytics [5], including possession and pace, offensive

and defensive ratings and the well-known Four Factors [8]. Useful graphical tools are also provided in order to visualize comparisons among teams and players on some selected statistics (bar-line plots, radial plots, scatterplots, bubble plots). In addition, indexes and graphs are included in the package for variability and inequality analyses. In addition, if the players' positions on the court are available, very nice shot charts can be created.

As an example of the wide range of basic analyses available in the package `BasketballAnalyzeR`, Figure 1 shows the radial plots representing the profiles of the players of Golden State Warriors who played at least 1200 minutes over the entire NBA championship 2017/2018, according to the following seven variables: 2-points shots made (P2M), 3-points shots made (P3M), free throws made (FTM), total rebounds per minute played (REB), assists per minute played (AST), steals per minute played (STL) and blocks per minute played (BLK). Each player is represented by a polygon that visually describes his profile according to the considered variables. In 1, players with different characteristics and different roles are considered. It would be also interesting to compare players playing the same role in different teams, and this is certainly possible with `BasketballAnalyzeR`.

Another example is in Figure 2, showing the shot chart of Stephen Curry. It highlights Curry's shot positions across the court: each point represents a shoot and is depicted in the player's position at the moment of shooting. In the top panel, points are colored according to whether the shot was made or missed. In the bottom panel, the court is split into a chosen number of sectors (here, five) that are colored according to a selected aggregate statistics (here, the average play length, in seconds, when the shot is attempted). Figure 2 shows that Stephen Curry attempts long distance shots and mid-range shots early in the play (prevalence of blue and green), while as the play length increases, he tends to shoot close-range shots.

4 Advanced analytics

Also advanced basketball analytics can be performed by `BasketballAnalyzeR`. First of all, it is possible to explore associations between variables, with a focus on visual tools, like maps displaying individual cases (teams or players) according to their similarity (based on multidimensional scaling) or graphical networks showing relationships among players. Also, shots density can be estimated (with respect to exogenous variables, such as the shot distance) and graphically represented. The package also allows to perform hierarchical and nonhierarchical cluster analysis, with a wide range of graphical tools able to depict similarities and differences among the obtained clusters that make it easy to characterize the clusters. Finally, some selected statistical models are included in the package (linear models and nonparametric regressions) and some specific issues are considered, such as the estimation of expected points or the scoring probability of one shot as a function of some variables (for example, time played or shot distance).

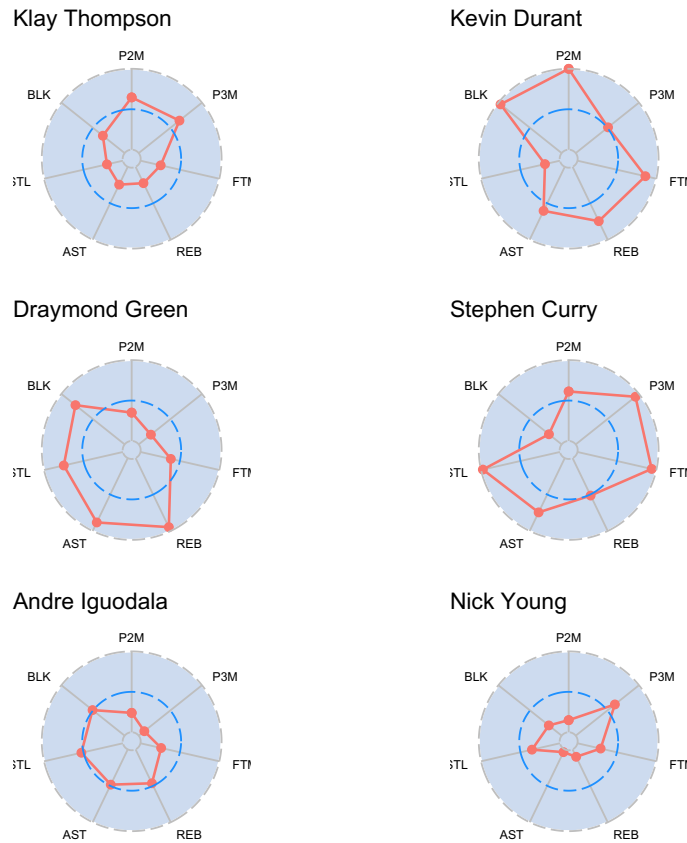


Fig. 1 Radial plots of the nine players of Golden State Warriors who played more than 1200 minutes in the NBA championship 2017/18. Standardized variables; Dashed blue line at zero

An example of advanced analytics is in Figure 3, which shows the correlation matrix concerning all those players who played at least 1000 minutes and the variables scored points (PTS), 3-point shots made (P3M), 2-point shots made (P2M), total rebounds per minute played (REB), assists per minute played (AST), turnovers per minute played (TOV), steals per minute played (STL) and blocks per minute played (BLK). The pairwise correlation coefficients, scatterplots, univariate histograms and boxplots are shown with a within-group analysis, that is by distinguishing players according to the Conference (East or West) in which their team plays.

As a second example, a hierarchical clustering has been performed in order to group NBA teams according to some game variables specifically created in order to account for both defensive and offensive abilities (including the famous Dean Oliver's four factors). Figure 4 shows the dendrogram (or cluster tree) obtained.

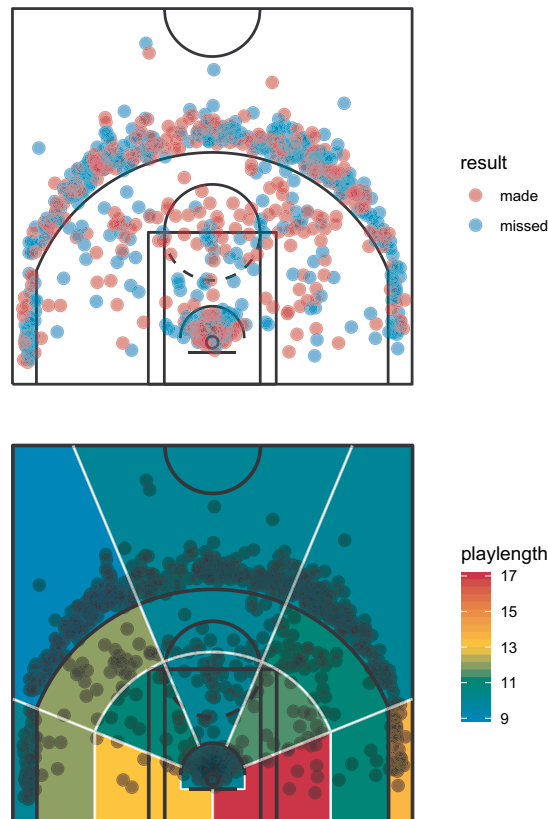


Fig. 2 Shot chart of Stephen Curry

5 Conclusions

Basketball analytics is growing fast, as documented by a wide literature, and the upcoming package of R `BasketballAnalyzerR` helps analysts to perform basic and advanced analyses on basketball data.

Although there is proof of the benefits of more advanced analysis beyond the success of teams, there are still coaches and legends of the sport who reject the practice of analytics. In our opinion, statistics cannot help basketball if it is intended as just points per game, number of assists and rebounds, and other basic aggregate results. Instead, statistical methods and models, together with the spread of a statistical culture, can help sport if modern approaches to basketball analytics are used and if analytics and technical experience are well integrated. To this end, it is important to talk with the experts, who pose the right questions to address by statistical analyses and give the right interpretations of the analytical solutions.

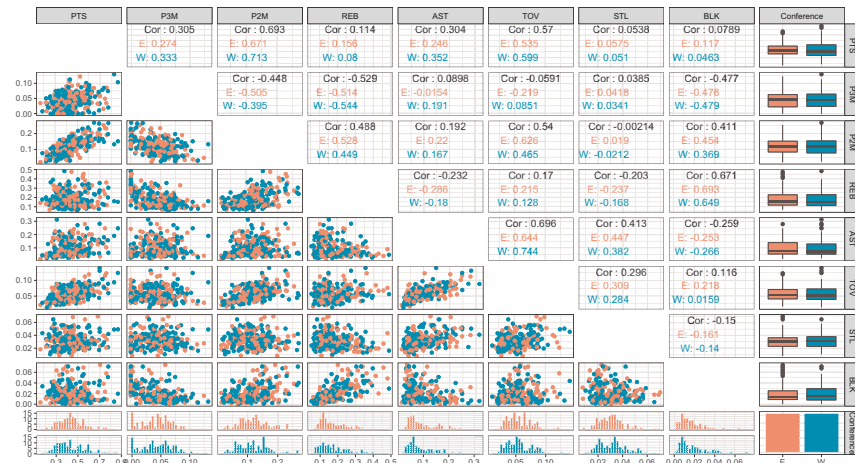


Fig. 3 Scatterplot matrix with within-group analysis (groups defined by Conference, East: E or West: W)

Acknowledgements Research carried out in collaboration with the Big & Open Data Innovation Laboratory (BODaI-Lab), University of Brescia (project nr. 03-2016, title “Big Data Analytics in Sports”, bdsports.unibs.it), granted by Fondazione Cariplo and Regione Lombardia.

References

1. Albert, J., Glickman, M.E., Swartz, T.B. and Koning, R.H.: Handbook of Statistical Methods and Analyses in Sports. CRC Press (2017)
2. Bianchi, F., Facchinetti, T., Zuccolotto, P.: Role revolution: towards a new meaning of positions in basketball. *Electronic Journal of Applied Statistical Analysis* **10**, 712–734 (2017)
3. Groll, A., Manisera, M., Schauburger, G., Zuccolotto, P.: Special Issue ‘Statistical Modelling for Sports Analytics’. *Statistical Modelling* **18**(5-6), (2018)
4. Groll, A., Manisera, M., Schauburger, G., Zuccolotto, P.: Special Issue ‘Statistical Modelling for Sports Analytics’. *Statistical Modelling* **19**(1), (2019)
5. Kubatko, J., Oliver, D., Pelton, K., Rosenbaum, D.T. and others: A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports* **3**, 1–22 (2007)
6. Manisera, M., Zuccolotto, P.: Basketball data science. CRC Press (Forthcoming)
7. Metulini, R., Manisera, M., Zuccolotto, P.: Modelling the dynamic pattern of surface area in basketball and its effects on team performance. *Journal of Quantitative Analysis in Sports* **14**, 117–130 (2018)
8. Oliver, D.: Basketball on paper: rules and tools for performance analysis. Potomac Books, Inc. (2004)
9. Sandri, M.: The R package `BasketballAnalyzer`. In: Manisera, M., Zuccolotto, P., Basketball data science, Chapter 6. CRC Press (Forthcoming)
10. Zuccolotto, P., Manisera, M., Kenett, R.S.: Special Issue ‘Statistics in Sports’. *Electronic Journal of Applied Statistical Analysis* **10**(3), (2017)
11. Zuccolotto, P., Manisera, M., Sandri, M.: Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions. *International Journal of Sports Science & Coaching* **13**, 569–589 (2018)

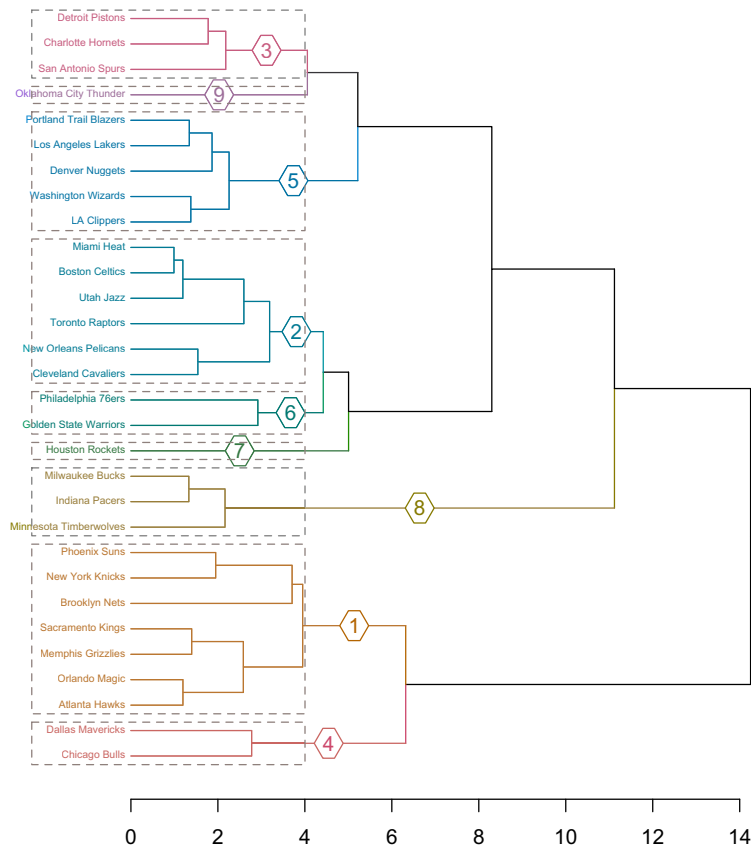


Fig. 4 Cluster tree of the hierarchical cluster analysis on the NBA teams based on selected game variables

Data Integration by Graphical Models

Utilizzo dei modelli grafici per l'integrazione dei dati

Daniela Marella and Paola Vicard and Vincenzina Vitale

Abstract Statistical matching aims at combining information obtained from different non-overlapping sample surveys. The main target is in constructing a complete synthetic data set where all the variables of interest are jointly observed. In this paper we propose the use of Bayesian Networks to deal with the statistical matching problem. Bayesian networks admit a recursive factorization of a joint distribution useful both for data integration and for evaluating the statistical matching uncertainty in the multivariate context.

Abstract Lo scopo dello statistical matching è stimare la distribuzione congiunta di variabili osservate in due campioni indipendenti, in cui la mancanza di osservazioni congiunte sulle variabili di interesse causa incertezza sul modello che ha generato i dati. Il lavoro propone l'utilizzo delle reti Bayesiane nell'ambito dello statistical matching. In particolare la fattorizzazione della distribuzione congiunta secondo il grafo diretto aciclico consente di ridurre la dimensione del problema con rilevanti vantaggi sia per l'integrazione dei dati che per l'analisi dell'incertezza in distribuzioni multivariate.

Daniela Marella

Dipartimento di Scienze della Formazione, Università Roma Tre, via del Castro Pretorio 20, 00185 Roma, e-mail: daniela.marella@uniroma3.it

Paola Vicard

Dipartimento di Economia, Via Silvio D'Amico 77, Università Roma Tre, 00145 Roma, e-mail: paola.vicard@uniroma3.it

Vincenzina Vitale

Dipartimento di Scienze Sociali ed Economiche, Sapienza Università di Roma, P.le Aldo Moro 5, 00185 Roma, e-mail: vincenzina.vitale@uniroma1.it

1 Introduction

The problem of statistical matching is becoming nowadays popular. Statistical matching is increasingly used in different contexts, such as economic analysis (see [12]), marketing analysis (as in [1]), confidentiality (see [17]). In practice, information usually comes from different databases, and there are not joint observations of all the characters of interest. The main target is in constructing a complete data set where all the variables of interest are jointly observed, see [6].

Formally, let (X, Y, Z) be a multivariate random variable (rv) with joint discrete distribution P . Without loss of generality, let $X = (X_1, \dots, X_H)$, $Y = (Y_1, \dots, Y_K)$ and $Z = (Z_1, \dots, Z_T)$ be vectors of rvs of dimension H , K , T , respectively. Statistical matching is defined as the estimation of (X, Y, Z) joint distribution (or of some of its parameters) when

1. Y is observed in a sample A of size n_A ;
2. Z is observed in a sample B of size n_B ;
3. A and B are independent, and the sets of observed units in the two samples do not overlap;
4. A set X of additional variables is observed both in A and B . Such common variables X are called matching variables.

Then, the units in A have missing Z values and the units in B have missing Y values. The joint distribution of (X, Y, Z) is not identifiable from samples A and B due to the lack of joint observations on Z and Y given X . Generally speaking, in order to overcome such a problem two approaches have been considered in the literature. The first approach uses techniques based on the conditional independence assumption between Y and Z given X (CIA assumption) see, e.g., [14]. The second approach uses techniques based on the external auxiliary information regarding the statistical relationship between Y and Z , e.g. an additional file C where (X, Y, Z) are jointly observed is available, as in [22].

However, it is possible that neither case is appropriate: the CIA assumption is usually a misspecified assumption as discussed in [21] and [19], while external auxiliary information is hardly ever available. Consequently, the third group of techniques addresses the so called *identification problem*. Roughly speaking, the lack of joint information on the variables of interest is the cause of uncertainty about the model of (X, Y, Z) since the sample information provided by A and B is actually unable to discriminate among a set of plausible models for (X, Y, Z) .

For instance, in a parametric setting and for $K = T = 1$, the consequence of the identifiability problem is that only ranges of plausible values on the missing records obtained from models fitted to the available sample information can be defined. Intervals defined by these ranges are known in the literature as uncertainty intervals. Uncertainty in statistical matching is analyzed in [10], [20], [13], [16], [7], [2], [3], [4] and [5].

In this paper we propose the use of graphical models and, in particular, of Bayesian networks (BNs) to deal with the statistical matching in the identification problem framework for categorical data. The first attempt for statistical matching

of discrete data by BNs is in [8] where, using the connection between conditional independence and d -separation criterion, the CIA is assumed. When the CIA model is not adequate, the application of the standard inferential procedures may result in highly misleading results. In section 2 the concept of uncertainty in statistical matching when BNs are used is discussed.

2 Uncertainty in Statistical Matching by Bayesian Networks

A Bayesian network (BN) is a probabilistic graphical model encoding a joint probability distribution by specifying the set of random variables and their conditional independence statements via a direct acyclic graph (DAG) and the set of conditional probability distributions associated to the nodes of the graph, see [15]. Then, the network consists of two components. The first component is a DAG. The second component is the set of all parameters in the network, that is the set of local probability distributions representing marginal and conditional probabilities.

The use of BNs in the statistical matching problem is motivated by the following advantages: (i) BNs can consider prior knowledge; (ii) BNs are widely used to describe dependencies among variables in multivariate distributions; (iii) BNs admit convenient recursive factorizations of their joint probability useful for data integration and uncertainty evaluation in a multivariate context.

As previously stressed, the lack of joint information on the variables of interest is the cause of uncertainty about the model of (X, Y, Z) then in the BNs approach both the DAG and its local distributions can be partially estimated from the information available in samples A and B , respectively. As a consequence, two kinds of uncertainty can be distinguished:

1. uncertainty regarding the association structure between the variables not jointly observed Y and Z ;
2. given the DAG, the uncertainty regarding the local probability distributions associated to the edges between Y and Z variables.

With regard to the uncertainty in the association structure, let P be the joint probability distribution of (X, Y, Z) associated to the DAG $G_{XYZ} = (V, E)$ consisting of a set of vertices V and a set E of directed edges between pairs of nodes. Let us denote by G_{XY} and G_{XZ} the DAGs estimated on sample A and B , respectively. As in [8] G_{XY} and G_{XZ} can be estimated subject to the condition that the association structure of the common variables X is fixed. Specifically, first of all, the DAG G_X is estimated on the overall sample $A \cup B$. Secondly, given G_X , we proceed to estimate the association structure between (X, Y) and (X, Z) on the basis of sample data in A and B , respectively.

In order to define the class of all plausible joint distributions for (X, Y, Z) the estimate collapsibility definition must be recalled. Formally, we say that the joint probability distribution P is estimate collapsible over Z_t if

$$\widehat{P}(X, Y, Z \setminus \{Z_t\}) = \widehat{P}_{G_{XYZ \setminus \{Z_t\}}}(X, Y, Z \setminus \{Z_t\}). \quad (1)$$

That is, the estimate $\widehat{P}(X, Y, Z \setminus \{Z_t\})$ of $P(X, Y, Z \setminus \{Z_t\})$ obtained by marginalizing the maximum likelihood estimate (MLE) of $\widehat{P}(X, Y, Z)$ under the original DAG model (G_{XYZ}, P) coincides with the MLE under the DAG model $(G_{XYZ \setminus \{Z_t\}}, P)$, see [11]. In terms of graphs a concept equivalent to estimate collapsibility is the c -removability, as discussed in [11].

Then, the class of plausible joint distributions for (X, Y, Z) can be defined as the set of all distributions satisfying the estimate collapsibility over Y and Z , respectively. It can be described as follows

$$\mathcal{P}_{XYZ} = \{P : \widehat{P}(X, Y) = \widehat{P}_{G_{XY}}(X, Y), \widehat{P}(X, Z) = \widehat{P}_{G_{XZ}}(X, Z)\} \quad (2)$$

or equivalently by using the graph of the model structure, the class can also be defined as the class of plausible DAGs G_{XYZ} where the variables Z and Y are c -removable, respectively. Formally

$$\mathcal{G}_{XYZ} = \{G_{XYZ} : Z \text{ is removable}, Y \text{ is removable}\} \quad (3)$$

Under the CIA, the class (2) is composed by a single joint probability distribution defined as $P(X, Y, Z) = P(X)P(Y|X)P(Z|X)$ and the class (3) collapses into a single graph given by $G_{XYZ}^{CIA} = G_{XY} \cup G_{XZ}$ where Y and Z are d -separated by the set X . The CIA network is always included in the class (3).

Note that, when the CIA does not hold, in order to choose a plausible DAG from the class \mathcal{G}_{XYZ} , it is important to have extra-sample information on the dependence structure that is generally available or can be elicited by experts.

With regard to the uncertainty in the local probability distributions associated to the edges between Y and Z variables, suppose that a DAG G_{XYZ}^* has been selected from the class of plausible DAGs \mathcal{G}_{XYZ} . Let P^* be the joint probability distribution associate to G_{XYZ}^* . According to G_{XYZ}^* the joint probability distribution P^* can be factorized into local probability distributions some of which can be estimated from the available sample information while other not due to the absence of joint observation on Y and Z variables. Uncertainty for categorical variables is dealt with in [7] where parameters estimation is performed according to the maximum likelihood principle.

Clearly, due to the missing data structure the parameter estimate which maximizes the likelihood function is not unique and the set of maximum likelihood estimates is called likelihood ridge. All the distributions in the likelihood ridge are equally informative given the data.

Suppose that from the factorization of P^* , the unique parameter that can not be estimated is the joint probability $P^*(X_h, Y_k, Z_t)$, where X_h , Y_k and Z_t are discrete rvs with I , J and L categories, respectively. Analogously to (2), as far as θ is concerned, one can only say that it lies in the following set:

$$\Theta_R = \{\theta^* : \sum_l \theta_{ijl}^* = \widehat{\theta}_{ij.}, \sum_j \theta_{ijl}^* = \widehat{\theta}_{i.l}, \theta_{ijl}^* \geq 0, \sum_{ijl} \theta_{ijl}^* = 1\} \quad (4)$$

where $\hat{\theta}_{ij}$ and $\hat{\theta}_{i,l}$ are the MLE estimates of (X_h, Y_k) and (X_h, Z_t) from samples A and B , respectively. For details, see [7].

In order to exclude some parameters in Θ_R reducing the overall parameter uncertainty, constraints characterizing the phenomenon under study can be introduced. Clearly, the amount of uncertainty reduction depends on the informative degree of the imposed constraints. These constraints can be defined in terms of structural zero ($\theta_{ijl}^* = 0$ for some (i, j, l)) and inequality constraints between pairs of distribution parameters ($\theta_{ijl}^* < \theta_{i'j'l'}^*$ for some $(i, j, l), (i', j', l')$). The likelihood function maximization problem when constraints are imposed may be solved through a modified EM algorithm, see [24]. In conclusion, in evaluating the quality of statistical matching procedures both the uncertainty regarding the association structure and the uncertainty regarding the local probability distributions should be considered in order to assess how reliable is the use of $(G_{XYZ}^*, \hat{\theta}^*)$ as a surrogate of the actual BN (G_{XYZ}, θ) .

References

1. Breur, T. (2011). Data analysis across various media: Data fusion, direct marketing, click-stream data and social media. *Journal of Direct, Data and Digital Marketing Practice*, 13, 95-105.
2. Conti, P.L., Marella, D., Scanu, M.: Uncertainty analysis in statistical matching. *Journal of Official Statistics*, 28, 69-88, (2012)
3. Conti P.L., Marella D., Scanu M.: Uncertainty Analysis for statistical matching of ordered categorical variables. *Computational Statistics & Data Analysis*, 68, 311-325, (2013)
4. Conti, P.L., Marella, D., Scanu, M.: How far from identifiability? A nonparametric approach to uncertainty in statistical matching under logical constraints. *Communication in Statistics: Theory and Methods*, (2013)
5. Conti, P.L., Marella, D., Scanu, M.: Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*. DOI:10.1080/01621459.2015.1112803, (2015)
6. D'Orazio, M., Di Zio, M., Scanu, M.: *Statistical Matching: Theory and Practice*. Chichester: Wiley, (2006)
7. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. *Journal of Official Statistics*, 22, 137-157, (2006)
8. Endres, E., Augustin, T.: Statistical matching of Discrete Data by Bayesian Networks. *JMLR: Workshop and Conference Proceedings*, 52, 159-170, (2016)
9. Geiger, D., Verma, T., Pearl, J.: Identifying Independence in Bayesian Networks. *Networks*, 20, 507-534, (1990)
10. Kadane, J.B.: Some statistical problems in merging data files. In *Compendium of tax research*, Department of Treasury, U.S. Government Printing Office, Washington D.C., 159-179 (Reprinted in 2001, *Journal of Official Statistics*, 17, 423-433), (1978).
11. Kim, S.H., Kim, S.H.: A Note on Collapsibility in DAG Models of Contingency Tables. *Scandinavian Journal of Statistics*, 33, 575-590, (2006)
12. Leulescu, A. and Agafitei, M. (2013). Statistical matching: a model based approach for data integration. Eurostat - Methodologies and Working papers, doi:10.2785/44822.
13. Moriarity, C., Scheuren, F.: Statistical Matching: A Paradigm of Assessing the Uncertainty in the Procedure. *Journal of Official Statistics*, 17, 407-422, (2001)
14. Okner, B.: Constructing a new data base from existing microdata sets: the 1966 merge file. *Annals of Economic and Social Measurement*, 1, 325-342, (1972)

15. Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, (1998)
16. Rässler, S.: *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer, New York, (2002)
17. Reiter, J. (2009). Using multiple imputation to integrate and disseminate confidential micro-data. *International Statistical Review*, 77, 179-195.
18. Ridder, G. and Moffitt, R.: The Econometrics of Data Combination. *Handbook of Econometrics*, J.J. Heckmann and E.E. Leamer (eds). vol. 6A. Amsterdam: Elsevier, (2007)
19. Rodgers, W. L.: An Evaluation of Statistical Matching. *Journal of Business and Economic Statistics*, 2 91-102, (1984)
20. Rubin, D.B.: Statistical Matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economics Statistics*, 4, 87-94, (1986)
21. Sims, C. A.: Comments on: "Constructing a New Data Base From Existing Microdata Sets: the 1966 Merge File", by B.A. Okner. *Annals of Economic and Social Measurements*, 1, 343-345, (1972).
22. Singh, A.C., Mantel, H., Kinack, M., Rowe, G.: Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. *Survey Methodology*, 19, 59-79, (1993).
23. Tsamardinos, I., Brown, L. E., Aliferis, C. F.: The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65, 31-78, (2006).
24. Winkler, W.E.: Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 274-279, (1993).

A two-part finite mixture quantile regression model for semi-continuous longitudinal data

Maruotti Antonello, Merlo Luca and Petrella Lea

Abstract This paper develops a two-part finite mixture quantile regression model for semi-continuous longitudinal data. The components of the finite mixture are associated with homogeneous individuals in the population sharing common values of the model parameters. The proposed methodology allows heterogeneity sources that influence the first level decision process, that is, the model for the binary response variable, to influence also the distribution of the positive outcomes. Estimation is carried out through an EM algorithm without parametric assumptions on the random effects distribution. A penalized version of the EM algorithm is also presented to tackle the problem of variable selection. The suggested modelling framework has been discussed using the extensively investigated RAND Health Insurance Experiment dataset in the random intercept case.

Abstract In questo paper sviluppiamo un modello a due parti mistura per dati longitudinali. Le componenti della mistura catturano cluster di individui che condividono caratteristiche simili. La metodologia adottata suppone che l'eterogeneità nella popolazione influenzi entrambe le due parti del modello, cioè la parte binaria e la parte positiva. La stima dei parametri è ottenuta tramite l'algoritmo EM senza formulare assunzioni parametriche sulla distribuzione degli effetti casuali. Viene proposta, inoltre, una estensione dell'algoritmo per la selezione delle variabili. L'analisi empirica si concentra sul dataset RAND Health Insurance Experiment utilizzando un modello ad intercetta casuale.

Maruotti Antonello

Centre for Innovation and Leadership in Health Sciences, University of Southampton
Department of Economic, Political Sciences and Modern Languages, LUMSA, Rome, Italy, e-mail: a.maruotti@lumsa.it

Merlo Luca

Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, e-mail: luca.merlo@uniroma1.it

Petrella Lea

MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano 9, e-mail: lea.petrella@uniroma1.it

Key words: Two-part model, Quantile regression, Random effect models, Longitudinal data, Variable selection, Healthcare expenditure

1 Introduction

Two-part models (TPM), also known as Hurdle models, involve a mixture distribution consisting in a mixing of a discrete point mass (with all mass at zero) and a discrete or continuous random variable. The TPM introduces modelling flexibility by allowing the zero and the positive values to be generated by two different processes. In its basic formulation, the TPM assumes independence between these two processes. Here the modelling framework is extended to model dependence between the spike-at-zero and the positive outcome accounting for potential heterogeneity in both processes. When dealing with longitudinal or hierarchical data, zero inflation may also occur when repeated data are analyzed. In addition, because measurements recorded on the same individual are likely correlated, the potential association between dependent observations should be taken into account in order to provide a correct inference. In such cases, random effect models have been proposed to accommodate for within-subject correlation and between subject heterogeneity, the latter accounting for zero inflation; see [5, 8]. Random effect models accommodate such source of random variation by considering unobserved heterogeneity in the model parameters via individual-specific random coefficients. With a parametric distribution for the random coefficients, one may use either a Monte Carlo EM algorithm or a Maximum Likelihood approach via Gaussian quadrature for parameters estimation. As an alternative, a non-parametric approach can be adopted in which the distribution of the random effects is left unspecified and approximated by using a discrete finite mixture distribution in order to prevent inconsistent parameter estimates due to misspecification of the underlying distribution. Within this scheme, the components of the finite mixture represent clusters of individuals that share homogeneous values of model parameters [2]. In the analysis of real data, applied researchers commonly focusing on estimating the conditional mean, might miss some underlying truth on the entire conditional distribution of the response variable of interest. To overcome those problems, in this paper we propose a quantile regression approach that allows estimating the full range of the conditional quantiles of the response variable. In addition, it is very common that a large number of candidate covariates available are included in the initial stage of modelling for the consideration of removing potential modelling biases. In order to gain in parsimony and to conduct a variable selection procedure, we propose a penalized version of the EM algorithm (PEM) adding the Least Absolute Shrinkage and Selecting Operator (LASSO) L_1 penalty term of [9]. Eventually, this paper can be considered as a unifying framework for the works of [3, 7, 1, 2] modelling time-constant random effects in a two-part finite mixture quantile model for longitudinal data. We examine the empirical behaviour of the proposed approach by the analysis of a sample taken from the RAND Health Insurance Experiment (RHIE).

2 Methodology

Let y_{it} , $i = 1, \dots, N$, $t = 1, \dots, T$ be a semi-continuous variable for unit i at time t and let $\mathbf{b}_i = (\mathbf{b}_{i0}, \mathbf{b}_{i1})$ be a time-constant, individual-specific, random effects vector having distribution $f_{\mathbf{b}}(\cdot)$ with support \mathcal{B} where $\mathbb{E}[\mathbf{b}_i] = 0$ is used for parameter identifiability. Let us assume that the random variable y_{it} has probability density function given by:

$$f(y_{it}) = p_{it}^{d_{it}} \left[(1 - p_{it}) g(h(y_{it}) | y_{it} > 0) \right]^{1-d_{it}} \quad (1)$$

with

$$d_{it} = I(y_{it} = 0), \quad p_{it} = \Pr(y_{it} = 0) = \Pr(d_{it} = 1)$$

where d_{it} denotes the occurrence variable for unit i at time t , $g(\cdot)$ is the density function for the positive outcome and $h(y_{it})$ denotes the intensity variable, with $h(\cdot)$ being a transformation function of y_{it} . Formally, the model is completed by defining the linear predictors for the binary and the positive parts of the model. The spike-at-zero process is governed by a binary logistic model such that:

$$\text{logit}(p_{it} | \mathbf{v}_{it}) = \mathbf{v}_{it}' \boldsymbol{\gamma} + \mathbf{c}_{it}' \mathbf{b}_{i0} \quad (2)$$

where $\mathbf{v}_{it} = (v_{it1}, \dots, v_{itm})$ is the m dimensional set of explanatory variables, $\boldsymbol{\gamma}$ its corresponding parameter vector and \mathbf{c}_{it} is a subset of covariates of \mathbf{v}_{it} . As mentioned in the Introduction, in this paper we are interested in modelling the truncated at zero part using the quantile regression approach. In particular, the positive part of the dependent variable can be written by exploiting the equivariance property of quantiles to monotone transformations: if $h(\cdot)$ is monotonic increasing on \mathbb{R}_+ , e.g. the logarithm, the quantiles of the transformed variable are the transformed quantiles of the original one: namely, for a given quantile $\tau \in (0, 1)$, $Q_{\tau}(\cdot | \mathbf{x}_{it}) = h(Q_{\tau}(\cdot | \mathbf{x}_{it}))$ where $Q_{\tau}(\cdot)$ is the quantile function. That is, conditionally on \mathbf{b}_{i1} , we assume that, after log-transforming the outcome variable, that is $\tilde{y}_{it} = \log(y_{it})$, the conditional density in (1) is:

$$g(\tilde{y}_{it} | y_{it} > 0, \mathbf{x}_{it}, \mathbf{b}_{i1}) = g_{it} = \frac{\tau(1-\tau)}{\sigma_{\tau}} \exp \left\{ -\rho_{\tau} \left(\frac{\tilde{y}_{it} - \mu_{it}}{\sigma_{\tau}} \right) \right\}, \quad (3)$$

where $\mathbf{x}_{it} = (x_{it1}, \dots, x_{its})$ represents a covariates s dimensional vector which may differs from \mathbf{v}_{it} , σ_{τ} is the scale parameter and $\rho_{\tau}(\cdot)$ denotes the quantile asymmetric loss function of [4]. Eq (3) represents an Asymmetric Laplace Distribution (ALD) discussed in [10] whose location parameter μ_{it} is defined by the linear model:

$$\mu_{it} = \mathbf{x}_{it}' \boldsymbol{\beta}_{\tau} + \mathbf{z}_{it}' \mathbf{b}_{i1} \quad (4)$$

where \mathbf{z}_{it} is a subset of covariates of \mathbf{x}_{it} . It is easy to see that (4) defines μ_{it} to be the τ -th conditional quantile function of the working variable \tilde{y}_{it} given $y_{it} > 0$ and \mathbf{x}_{it} .

As it is clear from (3) and (4), all parameters β_τ , σ_τ and \mathbf{b}_{i1} depend on the quantile level τ . Naturally, if the simple random intercept model is adopted, we have $\mathbf{z}_{it} \equiv 1$.

Parametric assumptions on the distribution of the random coefficients, $f_{\mathbf{b}}(\cdot)$, can be too restrictive and misspecification of the mixing distribution can lead to biased parameter estimates. In addition, an important disadvantage of this approach lies in the required computational effort. For these reasons, we may rely on Non-parametric Maximum Likelihood (NPML) estimation theory of [6]: if $f_{\mathbf{b}}(\cdot)$ is left unspecified, we approximate it by using a discrete distribution on $G < N$ locations $\mathbf{b}_k = (\mathbf{b}_{0k}, \mathbf{b}_{1k})$, with associated probabilities defined by $\pi_k = \Pr(\mathbf{b}_i = \mathbf{b}_k)$, $i = 1, \dots, N$ and $k = 1, \dots, G$. That is, $\mathbf{b}_i \sim \sum_{k=1}^G \pi_k \delta_{\mathbf{b}_k}$ where δ_θ is a one-point distribution putting a unit mass at θ . Because responses are assumed to be independent conditional on the random vector \mathbf{b}_k , in this case, the likelihood of the model has the form:

$$L(\Phi) = \prod_{i=1}^N \left\{ \sum_{k=1}^G \prod_{t=1}^T p_{it}^{d_{it}} [(1 - p_{it}) g_{itk}]^{1-d_{it}} \pi_k \right\}, \quad (5)$$

where, depending on the k -th component of the mixture, the spike-at-zero process is governed by the binary logistic model, $\text{logit}(p_{it} | \mathbf{v}_{it}) = \mathbf{v}_{it}' \gamma + \mathbf{c}_{it}' \mathbf{b}_{0k}$, and where g_{itk} is the ALD density in (3) with location parameter given by $\mu_{it} = \mathbf{x}_{it}' \beta_\tau + \mathbf{z}_{it}' \mathbf{b}_{1k}$, for $k = 1, \dots, G$.

Locations \mathbf{b}_k and corresponding masses π_k represent unknown parameters, as well as the unknown number of components G , which should be estimated along with other model parameters via selection model techniques. Here, the optimal number of components is based on penalized likelihood criteria such as the AIC and the BIC. As it can be easily noticed, (5) represents the likelihood function of a finite mixture of a quantile TPM with parameters vector $\Phi = \{\gamma, \beta_\tau, \mathbf{b}_1, \dots, \mathbf{b}_G, \sigma_\tau, \pi_1, \dots, \pi_G\}$.

2.1 Estimation

Given the finite mixture representation in (5), each unit i can be conceptualized as drawn from one of G distinct groups: we denote with w_{ik} a discrete latent variable indicating component membership, i.e. the indicator variable that is equal to 1 if the i -th unit belongs to the k -th component of the finite mixture, and 0 otherwise. We obtain maximum likelihood estimates using the EM algorithm by treating the hidden random variable as missing and representing the complete data set as $(y_{it}, v_{it}, x_{it}, w_{ik})$ for $i = 1, \dots, N$, $t = 1, \dots, T$ and $k = 1, \dots, G$. The log-likelihood for the complete data has the following form:

$$\ell_c(\Phi) = \sum_{i=1}^N \sum_{k=1}^G w_{ik} \left\{ \log \left(\prod_{t=1}^T p_{it}^{d_{it}} [(1 - p_{it}) g_{itk}]^{1-d_{it}} \right) + \log(\pi_k) \right\}. \quad (6)$$

In the E-step of the algorithm, the presence of the unobserved group-indicator is handled by taking the conditional expectation of w_{ik} given the observed data and

the parameter estimates at the r -th iteration $\hat{\Phi}^{(r)}$. We replace w_{ik} by its conditional expectation $\hat{w}_{ik}^{(r)}$. The quantity $\hat{w}_{ik}^{(r)}$ is the posterior probability that unit i comes from the k -th component of the mixture model. Subsequently, conditionally on the posterior probabilities $\hat{w}_{ik}^{(r)}$, the M-step solutions are updated using an iteratively weighted least squares (IWLS) algorithm via an appropriate weighted quantile regression for cross sectional. Standard error estimates for model parameters are obtained through non-parametric bootstrap.

With a large number of predictors, one often would like to determine a smaller subset of covariates that exhibit the strongest effects. We implement the variable selection method by maximizing the penalized complete data log-likelihood. Compared to the EM, the PEM algorithm leaves the E-step unchanged and modifies the M-step introducing a penalty function to achieve shrinkage. The log-likelihood for the complete data has the following form:

$$\ell_{pen}(\Phi|\lambda) = \ell_c(\Phi) - \lambda J(\beta_\tau), \quad (7)$$

where $J(\beta_\tau) = \|\beta_\tau\|_1$ is the convex LASSO penalty function of [9], $\ell_c(\Phi)$ has been defined in (6) and λ is a tuning parameter that regulates the strength of the penalization assigned to the coefficients in the model which is selected via cross-validation.

3 Application

The above-presented methodology has been applied to study the driving factors of medical expenditures using the data from the RHIE. The RHIE experiment, conducted from 1974 to 1982, collects data from about 8000 enrollees in 2823 families, from six sites across the United States. It assesses how medical care costs affect a patient's use of health services and quality and it is regarded as the basis of the most reliable estimates of price sensitivity of demand for medical services. We consider one measure of utilization: the total spending on health services defined as the sum of outpatient, inpatient, drugs, supplies and psychotherapy expenses. Table 1 reports the number of free model parameters, the log-likelihood and the penalized likelihood criteria (AIC and BIC) for different number of mixture components G at three quantile levels $\tau = (0.25, 0.50, 0.75)$.

References

- [1] Alfò, M. and Maruotti, A. [2010], 'Two-part regression models for longitudinal zero-inflated count data', *Canadian Journal of Statistics* **38**(2), 197–216.
- [2] Alfò, M., Salvati, N. and Ranalli, M. G. [2017], 'Finite mixtures of quantile and m-quantile regression models', *Statistics and Computing* **27**(2), 547–570.

- [3] Geraci, M. and Bottai, M. [2006], ‘Quantile regression for longitudinal data using the asymmetric Laplace distribution’, *Biostatistics* **8**(1), 140–154.
- [4] Koenker, R. and Bassett, G. [1978], ‘Regression quantiles’, *Econometrica: Journal of the Econometric Society* **46**(1), 33–50.
- [5] Lam, K., Xue, H. and Bun Cheung, Y. [2006], ‘Semiparametric analysis of zero-inflated count data’, *Biometrics* **62**(4), 996–1003.
- [6] Lindsay, B. G. et al. [1983], ‘The geometry of mixture likelihoods: a general theory’, *The annals of statistics* **11**(1), 86–94.
- [7] Maruotti, A., Raponi, V. and Lagona, F. [2016], ‘Handling endogeneity and nonnegativity in correlated random effects models: Evidence from ambulatory expenditure’, *Biometrical Journal* **58**(2), 280–302.
- [8] Min, Y. and Agresti, A. [2005], ‘Random effect models for repeated measures of zero-inflated count data’, *Statistical modelling* **5**(1), 1–19.
- [9] Tibshirani, R. [1996], ‘Regression shrinkage and selection via the lasso’, *Journal of the Royal Statistical Society. Series B* **58**, 267–288.
- [10] Yu, K. and Moyeed, R. A. [2001], ‘Bayesian quantile regression’, *Statistics & Probability Letters* **54**(4), 437–447.

G	2	3	4	5	6
# par.	34	37	40	43	46
$\tau = 0.25$					
log-lik	-36157.85	-35818.09	-35691.69	-35615.53	-35566.29
AIC	72383.71	71710.18	71463.38	71317.06	71224.58
BIC	72610.99	71957.52	71730.77	71604.50	71532.08
$\tau = 0.50$					
log-lik	-35703.81	-35374.62	-35254.80	-35174.27	-35147.60
AIC	71475.63	70823.25	70589.60	70434.53	70387.19
BIC	71702.91	71070.58	70856.99	70721.98	70694.69
$\tau = 0.75$					
log-lik	-37379.27	-36614.51	-36422.58	-36374.00	-36313.25
AIC	74826.54	73303.02	72925.17	72834.00	72718.50
BIC	75053.82	73550.36	73192.56	73121.44	73026.00

Table 1 Number of parameters, log-likelihood and penalized likelihood criteria for different number of mixture components G at quantile levels $\tau = (0.25, 0.50, 0.75)$. The optimal number of components are displayed in boldface.

Multivariate change-point analysis for climate time series

Analisi di change-point multivariati per serie storiche climatiche

Gianluca Mastrantonio, Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio and Carlo Blasi

Abstract The aim of this work is to find individual and joint change-points in a large multivariate database of climate data. We model monthly values of precipitation, minimum and maximum temperature recorded in 360 stations covering all Italy for 60 years (12×60 months). The proposed three variate Gaussian change-point model exploits the Hierarchical Dirichlet process, allowing for a formalization that lets us estimate a different change-point model for each station. As stations possibly share some of the parameters of the trivariate normal emission distribution, this model framework provides an original definition of the change-points corresponding to changes in any subset of the 12 model parameters. In this paper, results for two stations in Southern Italy are shown as an example.

Abstract *L'obiettivo di questo lavoro è l'individuazione di change-point semplici e congiunti in un grande database multivariato di dati climatici: osservazioni mensili della precipitazione e delle temperature minima e massima registrate per 60 anni (12×60 mesi) in 360 stazioni dislocate su tutto il territorio nazionale. Proponiamo un modello Gaussiano trivariato basato sul processo di Dirichlet gerarchico che permette di stimare un modello di change-point diverso per ciascuna stazione. Poiché stazioni diverse possono condividere i valori di alcuni dei 12 parametri della distribuzione normale trivariata della risposta, questo modello permette di pervenire ad una definizione originale dei change-point in corrispondenza di cam-*

Gianluca Mastrantonio
Department of Mathematical Sciences, Politecnico di Torino

Giovanna Jona Lasinio
Department of Statistical Sciences, Sapienza Università di Roma

Alessio Pollice
Department of Economics and Finance, Università di Bari Aldo Moro

Giulia Capotorti, Carlo Blasi
Department of Environmental Biology, Sapienza Università di Roma

Lorenzo Teodonio
ICRCPAL, Ministry of Cultural Heritage and Activities and Tourism, Roma

biamenti in qualsiasi sottoinsieme di detti parametri. A titolo di esempio riportiamo i risultati della stima del modello per due stazioni nel Sud Italia.

Key words: Change-point model, Hierarchical Dirichlet process, Climate data

1 Introduction

Climate elements and regimes, such as temperature, precipitation and their annual cycles, primarily affect the type and distribution of plants, animals, and soils as well as their combination in complex ecosystems [2, 9]. The ecological classification of climate represents one of the basic steps for the definition and mapping of ecoregions, i.e. of broad ecosystems occurring in discrete geographical areas [1, 7]. In keeping with these assumptions, a hierarchical classification of Italian ecoregions was recently obtained by combining climatic diagnostic features with distribution patterns of biological diversity and physical characteristics of the environment [3]. The Italian ecoregions are arranged into four hierarchically nested tiers, which consist of 2 Divisions, 7 Provinces, 11 Sections and 35 Subsections. The climatic features adopted for the diagnosis and description of the Italian ecoregions refer to thermo-pluviometric data and bioclimatic indices that date back to the period 1955-2000.

From a botanical perspective, the analysis of change-points allows to detect abrupt changes in the climatic behaviour and supports inferences of the potential effects of these changes on ecosystem composition, functionality, distribution, and dynamics at different spatial and time scales [6, 5].

In order to simplify the joint distribution modeling of the climate variables, we standardize the temperatures and rescale the precipitation with its standard deviation; the latter is then seen as the realization of a latent variable belonging to the real line (\mathbb{R}) [8]. At each station the trivariate time series can then be assumed to come from a change-point (CP) model with multivariate normal emission distribution, parametrized using 12 parameters: 3 intercepts, 3 regressive coefficients for the time covariate, 3 variances and 3 correlations. We define the CP model using a modified version of the hierarchical Dirichlet process [11], which allows that different time series can share a subset of the 12 parameters and that some or all of these parameters change at each change-point.

2 The data

We consider monthly records of precipitation and min/max temperature at 360 monitoring stations over 60 years (1951-2010). The data were mostly obtained from National Institutions (ISPRA, CRA/CREA, Meteomont and ENEA) and local authorities [8]. Monthly records were obtained considering monthly cumulative pre-

precipitations and monthly averages of daily minimum and maximum air temperatures. Almost all original time series are affected by variable amounts of missing data. The full database reports $360 \times 60 \times 12$ entries.

Let us denote with $Y_1^*(\mathbf{s}, t)$, $Y_2^*(\mathbf{s}, t)$ and $Y_3^*(\mathbf{s}, t)$ the observed precipitation, minimum and maximum temperature, respectively, observed over the $n_s = 360$ stations for $n_t = 720$ months, starting from January 1950, where $\mathbf{s} \in \mathcal{S} \subset \mathbb{R}^2$ is a generic couple of geographic coordinates while $t \in \mathcal{T} \subset \mathbb{R}$, s is assumed to vary continuously in its domain, while time is discrete. Following [8], to facilitate model implementation we work with the following (latent) variables:

$$\begin{cases} Y_1(\mathbf{s}, t) = \frac{Y_1^*(\mathbf{s}, t)}{\sqrt{\text{Var}(Y_1^*(\mathbf{s}, t))}} & \text{if } Y_1^*(\mathbf{s}, t) > 0, \\ Y_1(\mathbf{s}, t) \leq 0 & \text{if } Y_1^*(\mathbf{s}, t) = 0, \end{cases}$$

$$Y_2(\mathbf{s}, t) = \frac{Y_2^*(\mathbf{s}, t) - \text{E}(Y_2^*(\mathbf{s}, t))}{\sqrt{\text{Var}(Y_2^*(\mathbf{s}, t))}},$$

$$Y_3(\mathbf{s}, t) = \frac{Y_3^*(\mathbf{s}, t) - \text{E}(Y_3^*(\mathbf{s}, t))}{\sqrt{\text{Var}(Y_3^*(\mathbf{s}, t))}},$$

In particular precipitations are modeled through a latent variable Y_1 that is marginally Gaussian and its negative tail is associated to the event “no precipitation” [8].

3 The model

Our main idea is to model each trivariate time series $Y(\mathbf{s}) = \{Y_1(\mathbf{s}, t), Y_2(\mathbf{s}, t), Y_3(\mathbf{s}, t)\}_{t \in \mathcal{T}}$ using a CP model with a multivariate normal emission distribution and the following features:

$$f(\mathbf{y}|\{\beta_{s,k}^*, \Sigma_{s,k}^*\}_{k \in \mathbb{Z}}) = \prod_{s \in \mathcal{S}} \prod_{t \in \mathcal{T}} \phi_3(\mathbf{y}_{t,s} | X_t \beta_{t,s}, \Sigma_{t,s}), \quad (1)$$

$$\beta_{t,s}, \Sigma_{t,s} | (\beta_{t-1,s}, \Sigma_{t-1,s}) = (\beta_{s,k}^*, \Sigma_{s,k}^*) \sim G_s^k, \quad (2)$$

$$G_s^k = \sum_{j \in \mathbb{Z} \setminus \{1, \dots, k-1\}} \frac{\pi_{s,j}}{1 - \sum_{h=1}^{k-1} \pi_{s,h}} \delta(\beta_{s,j}^*, \Sigma_{s,j}^*),$$

$$G_s = \sum_{k \in \mathbb{Z}} \pi_{s,k} \delta(\beta_{s,k}^*, \Sigma_{s,k}^*) \sim \text{Dir}(\alpha, G_0), \quad (3)$$

$$G_0 = \left(\prod_{i=1}^3 \prod_{j=1}^2 G_{\beta_{ij}^*} \right) \left(\prod_{i=1}^3 G_{\sigma_i^{2,*}} \right) \left(\prod_{i=1}^2 \prod_{j=i+1}^3 G_{\rho_{ij}^*} \right), \quad (4)$$

$$G_{\beta_{11}^*} = \sum_{k \in \mathbb{Z}} \xi_{\beta_{11,k}^*} \delta(\beta_{11,k}^*) \sim \text{Dir}(\gamma, H_{\beta_{11}^*}), \quad (5)$$

...

$$G_{\rho_{23}^*} = \sum_{k \in \mathbb{Z}} \xi_{\rho_{23,k}^*} \delta(\rho_{23,k}^*) \sim \text{Dir}(\gamma, H_{\rho_{23}^*}), \quad (6)$$

where k indexes the time series regimes, \mathbf{X}_t is a 1×2 vector with 1 as first element and t as second, $\phi_3(\cdot)$ is a trivariate normal density and δ is the indicator function.

The trivariate normal density in (1) is parametrized using 12 parameters: 6 regressive coefficients (β), 3 variances σ^2 and 3 correlations ρ . For each of them we have a corresponding distribution drawn from a Dirichlet process, in (5)-(6). All these distributions are combined to obtain the discrete distribution G_0 in (4), with atoms given by (5)-(6) and weights obtained as the products of the ones of (5)-(6).

As in the standard Hierarchical Dirichlet process, a common Dirichlet based distribution, here G_0 , is used as base distribution for DP draws, see (3), where the index s recalls that we choose a distribution for each observed location (station). The G_s distributions share the same set of atoms with different weights.

Equations (1) -(2) define CP models for each station. At each time t , the values of the regressive coefficient and covariance matrix of the normal emission distribution are drawn from the discrete distribution G_s^k which allows them to assume the same value of the previous time or a new one that has never been observed previously at the specific location s . This is achieved by defining G_s^k as G_s removing the atom corresponding to the already observed values and with rescaling the weights to sum to 1. Due to the way we define the base measure G_0 , the atoms of G_s^k may share some parameters. In other words, the occurrence of a change point does not imply that all 12 parameters change. The minimal requirement for a CP detection is that at least one change in the parameter set occurs. Further, G_s and $G_{s'}$ have the same set of atoms, implying that the two time series can have all or some of parameters in common.

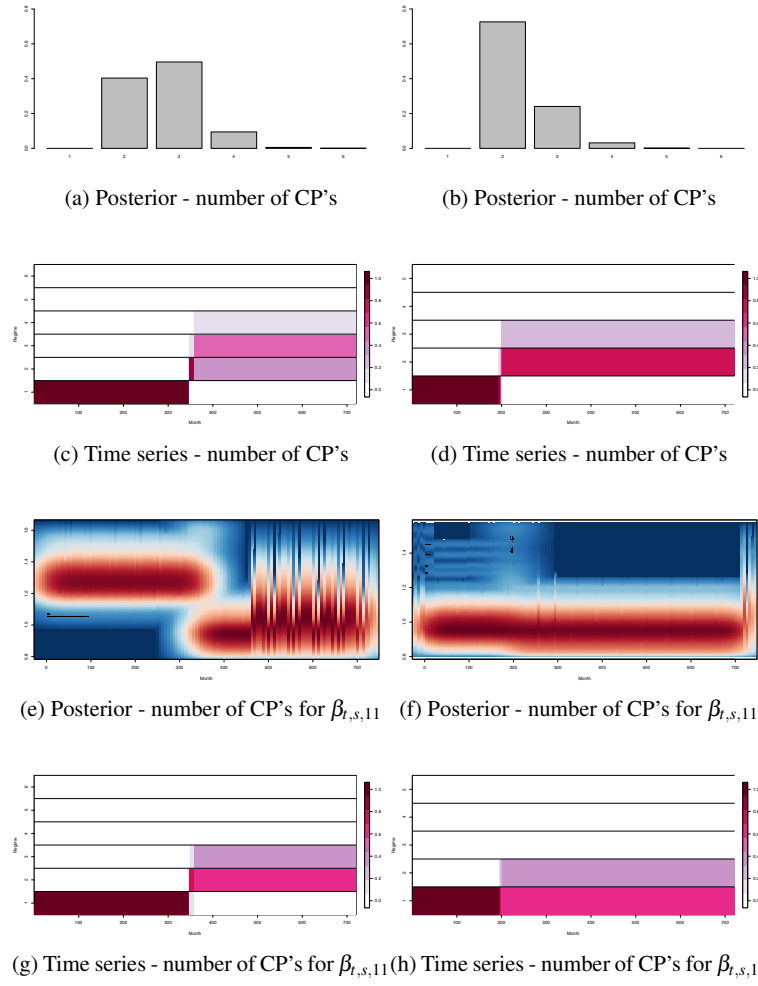


Fig. 1: Posterior distributions and time-series of the number of change-points for all parameters (top: a, b, c, d) and for $\beta_{t,s,11}$ (bottom: e, f, g, h) for the stations *Lucera* (left; a, c, e, g) and *Vieste* (right: b, d, f, h).

4 Some Results

In this section we show some of the inferences that can be obtained from the estimation of the proposed model. We describe part of the results for the stations *Lucera* and *Vieste* both located in the Puglia region and belonging to a small ecoregion. From Figures 1 (a) and (b) we see that the maximum a posteriori (MAP) number of change-points in Lucera is three, two CP's have high probability too, while in Vieste

we observe two CP's. Figures 1 (c) and (d) show the probability that at a given time (horizontal) the time series is following a specific regime (vertical), the darker the color the largest the probability. We can observe that even if Lucera (c) registers three possible regimes, the second one lasts less than one year. In Figures 1 (g) and (h) we see the time series of regimes switching for the intercept of the precipitation. It is interesting to note that it has MAP number of change-points equal to two for Lucera while it is just one for Vieste (no change-points). Vieste change points are due to a change in temperature regime (not shown here), since both increase in mean level. To conclude, Figures 1 (e) and (f) show the posterior densities of the considered parameter for each time point. The time points are represented by the x-axis, in the y-axis there are the values assumed by the parameter and the color represents the density, with blue equals to zero and the darkest red the maximum value. We can see a clear change of value at point 320, for the precipitation at Lucera while no strong evidence of change-point is obtained for the precipitation in Vieste. Notice that, starting from time 320 (when Lucera's change-point is found), there is a probability of 0.7 that the intercept of the precipitation has the same value in the two stations, as we can see from Figures 1 (e) and (f).

5 Conclusions

In this work we present the preliminary results of the finding of joint and individual change-points in a large database of climate data. We model monthly values of precipitation, minimum and maximum temperature recorded all over Italy for 60 years. The model was implemented on the TeraStat cluster [4]. The code is written in R/C++, and uses the openMP library [10] to perform parallel computing. Our proposal allows for a very rich inference on the joint CP detection. Posterior estimates are obtained in 2 days, with 40000 iterations per day and 10 GB of ram usage. Initial results are very encouraging and are partially limited only by computational issues, that we plan to solve in the near future.

Acknowledgements The work of the first three authors is partially developed under the PRIN2015 supported-project Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTat) funded by MIUR (Italian Ministry of Education, University and Scientific Research) (20154X8K23-SH3).

References

1. Bailey, R.G.: Delineation of ecosystem regions. *Environmental Management* **7**(4), 365–373 (1983)
2. Bailey, R.G.: Identifying ecoregion boundaries. *Environmental Management* **34**(Suppl 1), S14–S26 (2004). DOI <https://doi.org/10.1007/s00267-003-0163-6>

3. Blasi, C., Capotorti, G., Copiz, R., Guida, D., Mollo, B., Smiraglia, D., Zavattero, L.: Classification and mapping of the ecoregions of Italy. *Plant biosystems - An International Journal Dealing with all Aspects of Plant Biology* **148**(6), 1255–1345 (2014). DOI 10.1080/11263504.2014.985756. URL <http://dx.doi.org/10.1080/11263504.2014.985756>
4. Ferraro Petrillo, U., Raimato, G.: Terastat computer cluster for high performance computing. “<http://www.dss.uniroma1.it/en/node/6554>” Department of Statistical Science Sapienza university of Rome (2014)
5. Kong, D., Zhang, Q., Singh, V.P., Shi, P.: Seasonal vegetation response to climate change in the northern hemisphere (1982–2013). *Global and planetary change* **148**, 1–8 (2017)
6. Liu, Y., Lei, H.: Responses of natural vegetation dynamics to climate drivers in China from 1982 to 2011. *Remote Sensing* **7**, 10,243–10,268 (2015)
7. Loveland, T.R., Merchant, J.M.: Ecoregions and ecoregionalization: geographical and ecological perspectives. *Environmental Management* **34**((Suppl 1): S1) (2004). DOI 10.1007/s00267-003-5181-x
8. Mastrantonio, G., Jona Lasinio, G., Pollice, A., Capotorti, G., Teodonio, L., Genova, G., Blasi, C.: A hierarchical multivariate spatio-temporal model for clustered climate data with annual cycles. *Annals of Applied Statistics* (in press)
9. Metzger, M.J., Bunce, R.G.H., Jongman, R.H.G., Sayre, R., Trabucco, A., Zomer, R.: A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring. *Global Ecology and Biogeography* **22**(5), 630–638 (2013). DOI 10.1111/geb.12022. URL <http://dx.doi.org/10.1111/geb.12022>
10. OpenMP Architecture Review Board: OpenMP application program interface version 3.0 (2008). URL <http://www.openmp.org/mp-documents/spec30.pdf>
11. Teh, Y.W., Jordan, M.I.: Hierarchical Bayesian nonparametric models with applications. In: N. Hjort, C. Holmes, P. Müller, S. Walker (eds.) *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press (2010)

A divide-et-impera approach for the spatial prediction of object data over complex regions

Un approccio divide-et-impera per la previsione spaziale di dati oggetto su regioni complesse

Alessandra Menafoglio e Piercesare Secchi

Abstract We consider the problem of performing a spatial statistical analysis of complex data when a global model is untractable (e.g., because of the structure of the spatial domain) or not appropriate to describe their variability (e.g., in case of strong non-stationarities). We shall illustrate a novel system of ideas, based on Random Domain Decompositions, that allows to deal with the problem complexity through a *divide-et-impera* approach. The methodology we propose embraces the viewpoint of Object Oriented Spatial Statistics (O2S2, Menafoglio and Secchi (2017)), and allows dealing with general types of data (functional data, distributional data, Riemannian data). The approach shall be illustrated on a case study aimed at the spatial analysis of dissolved oxygen depletion in the Chesapeake Bay (US).

Abstract Si considera il problema della analisi statistica spaziale di dati complessi, qualora un modello globale sia intrattabile (e.g., per la struttura del dominio spaziale) o non appropriato a descrivere la loro variabilità (e.g., per via di forti non-stazionarietà). Sarà illustrato un nuovo sistema di idee, basato su decomposizioni aleatorie del dominio, che consente di trattare la complessità del problema attraverso un approccio divide-et-impera. La metodologia proposta abbraccia il punto di vista della Object Oriented Spatial Statistics (O2S2, Menafoglio and Secchi (2017)), e consente di analizzare tipi generali di dati (dati funzionali, distribuzionali o Riemanniani). L'approccio sarà illustrato tramite un caso studio finalizzato all'analisi spaziale della riduzione dell'ossigeno disciolto nella Chesapeake Bay (US).

Key words: Object Oriented Data Analysis, Local Stationarity, Bagging algorithm

Alessandra Menafoglio

MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano,
e-mail: alessandra.menafoglio@polimi.it

Piercesare Secchi

MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano,
e-mail: piercesare.secchi@polimi.it

The analysis of complex data distributed over large or highly textured regions poses new challenges for spatial statistics. Although methods to deal with spatial object data have been successfully applied in several environmental studies (e.g., Menafoglio et al. (2013, 2014, 2016b,a)), they rely upon global models for the spatial dependence of the field, that are rarely usable in the presence of large, textured or convoluted domains, with holes or barriers.

In this communication, we focus on a novel system of ideas based on Random Domain Decompositions, that allows to decompose a problem whose complexity is untractable at a global scale, in a (random) system of local tractable problems, in a divide-et-impera framework. We shall illustrate our recent methodology presented in (Menafoglio et al., 2018), that enables one to perform spatial prediction of object data distributed in highly texture regions. In this framework, we propose to perform repeated Random Domain Decompositions (RDDs) of the study area, each defining a set of homogeneous sub-regions where to perform local object-oriented spatial analyses, under stationarity assumptions. This system of *weak* analyses are then aggregated into a final global analysis, in a *bagging* setting (Breiman, 1996). In this broad framework, the complexity of the domain can be taken into account by defining its partitions through a non-Euclidean metric that properly represents the adjacency relationships among the observations over the domain.

The method we propose is entirely general, and prone to be used with numerous types of object data (e.g., functional data, density data or manifold data), being grounded upon the theory of Object Oriented Spatial Statistics (O2S2, Menafoglio and Secchi (2017)). In this vein, we shall first describe the case of functional data embedded into a Hilbert space, whose geometry allows to use linear geostatistical methods in each subregion defined by the RDD. We shall finally provide insights on a recent extension of the method that allows for the analysis of data belonging to a Riemannian (Menafoglio et al., 2018). Here, the RDD may not only be used to better describe the adjacency relation among data when these are distributed a textured domain, but also to provide a system of linear approximations (tangent spaces) of the Riemannian manifold and allow for the application of linear geostatistical methods in this case too.

As an insightful illustration of the potential of the methodology, we shall consider the spatial prediction of aquatic variables in estuarine systems, that are non-convex and very irregularly shaped regions where the narrow areas of land between adjacent tributaries act as barriers. Here, we focus on the analysis and spatial prediction of distributional data (density functions and covariance matrices) relevant to the study of dissolved oxygen depletion in the Chesapeake Bay (US).

References

- Breiman, L. (1996). Bagging predictors. *Machine Learning* 24, 123–149.
 Menafoglio, A., A. Guadagnini, and P. Secchi (2014). A Kriging approach based on Aitchison geometry for the characterization of particle-size curves in heteroge-

- neous aquifers. *Stochastic Environmental Research and Risk Assessment* 28(7), 1835–1851.
- Menafoglio, A., A. Guadagnini, and P. Secchi (2016a). Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a bayes space approach. *Water Resources Research* 52(8), 5708–5726.
- Menafoglio, A., D. Pigoli and P. Secchi (2018). Kriging Riemannian Data via Random Domain Decompositions *MOX-report 64/18*, Politecnico di Milano.
- Menafoglio, A. and P. Secchi (2017). Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. *European Journal of Operational Research* 258(2), 401–410.
- Menafoglio, A., P. Secchi, and M. Dalla Rosa (2013). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics* 7, 2209–2240.
- Menafoglio, A., P. Secchi, and A. Guadagnini (2016b). A Class-Kriging Predictor for Functional Compositions with Application to Particle-Size Curves in Heterogeneous Aquifers. *Mathematical Geosciences* 48, 463–485.
- Menafoglio, A., G. Gaetani, and P. Secchi (2018). Random Domain Decompositions for object-oriented Kriging over complex domains. *Stochastic Environmental Research and Risk Assessment* 32(12), 3421–3437.

A strategy for the matching of mobile phone signals with census data

Una strategia per l'abbinamento di segnali di telefonia mobile con dati censuari

Rodolfo Metulini and Maurizio Carpita

Abstract Administrative data allows us to count for the number of residents. The geo-localization of people by mobile phone, by quantifying the number of people at a given moment in time, enriches the amount of useful information for “smart” (cities) evaluations. However, using Telecom Italia Mobile (TIM) data, we are able to characterize the spatio-temporal dynamic of the presences in the city of just TIM users. A strategy to estimate total presences is needed. In this paper we propose a strategy to extrapolate the number of total people by using TIM data only. To do so, we apply a spatial record linkage of mobile phone data with administrative archives using the number of residents at the level of “sezione di censimento”.

Abstract *I dati amministrati ci permettono di avere l'informazione sul numero di residenti. La geolocalizzazione delle persone tramite telefoni cellulari permette di quantificare il numero di persone presenti in ciascun istante temporale e arricchisce la mole di informazioni utilizzabili nell'ambito delle valutazioni di “smart city”. Tuttavia, utilizzando i dati di Telecom Italia Mobile (TIM) e' possibile caratterizzare le dinamica spazio temporale delle presenze dei soli utenti TIM. Una strategia per stimare il numero totale di presenze e' necessaria. In questo lavoro proponiamo una strategia per ottenere la dinamica complessiva delle presenze partendo dalla disponibilita' dei soli dati TIM. Per fare cio', applichiamo un record linkage spaziale tra i dati telefonici e il numero di residenti ad un livello di “sezioni di censimento”.*

Key words: Record Linkage; Big Data; Smart Cities; Mobile Phone; Spatial Analysis

Rodolfo Metulini

Department of Economic and Management, University of Brescia, Contrada Santa Chiara, 50, 25122 Brescia, e-mail: rodolfo.metulini@unibs.it

Maurizio Carpita

Department of Economic and Management, University of Brescia, Contrada Santa Chiara, 50, 25122 Brescia, e-mail: maurizio.carpita@unibs.it

1 Introduction

A crucial aspect for the well-being of an urban area in the context of the much discussed “Smart-City” argument refers to the monitoring of the dynamic of people’s presences. To do so, thanks to a research collaboration of DMS StatLab (<https://sites.google.com/a/unibs.it/dms-statlab/>) with the Statistical Office of the Municipality of Brescia, we use mobile phone data provided by Telecom Italia Mobile (TIM), quantifying the number of TIM users connected to the smartphone over a spatial grid characterized by its latitude and longitude, and over the time.

We estimate the presence of TIM users in the city of Brescia by classifying similar days in terms of both the spatial and the temporal dimension. In doing so, we use a larger (≈ 2 years) dataset compared to that used in similar works (Zanini et al. [1] for example). To manage with high dimensional data, we employ a multi-stage procedure that, in the first step, converts the data matrix containing the values of the grid (2-D) at each point in time to a vector (1-D) of features using a method borrowed from image processing (the Histogram of Oriented Gradients - HOG). The procedure, by classifying similar days using a mix of traditional (k-means) and functional data clustering techniques, is able to identify the dynamic related to the presences of TIM users in the city.

However, for a proper evaluation, all the people should be counted in. In absence of the data about other companies’ users, the coefficient related to TIM market share could be applied. Unfortunately, this percentage is just available at country level, while we have reason to think that the market share varies along different cities due to socio-economic and demographic reasons. So, the aim of this work focuses on the estimation of the number of people at a city level. In doing so, we employ a record linkage based on comparing residents from administrative archives with the number of TIM users on specific areas of the grid during a sample of proper time periods.

All in all, this work regards the match of administrative archives with data from technological devices, adopting a method that reminds spatial record linkages (Blum and Calvo [2] and Xu et al. [3]), for the purpose of quantifying the presences in the city when some data are missings.

Section 2 describes the mobile phone data, section 3 summarizes the procedure adopted to classify similar days and to quantify the number of TIM users in those days. Section 4 outlines the record linkage strategy adopted and section 5 discusses and concludes.

2 Data

This work focuses on mobile phone data provided by Telecom Italia Mobile (TIM), which is currently the largest operator in Italy in this sector. These data sometimes goes with the name of “Erlang” measures. In detail, the data refer to the mobile phone activity recorded in the period April 1st 2014 to August 11th 2016, in a rect-

angular region defined by latitude 45.21° N - 46.36° N and longitude 9.83° N - 10.85° N. Data were aggregated into 923×607 cells of 150 m^2 size each. Data are available at intervals of 15 minutes, for a total of more than 40,000 millions of records collected. For each cell and for each time interval, the corresponding record refers to the average number of mobile phones simultaneously connected to the network in that area in that time interval. The mobility feature of these data is hidden, in the sense it is not possible to trace the single person over the time.¹ Similar data has been used by Carpita and Simonetto [5], who analyzed the presence of people during big events in the city of Brescia, by Zanini et al. [1], who find, by mean of a Independent Component Analysis (ICA), a number of spatial components that separate main areas of the city of Milano, and by others (Metulini and Carpita [6] Manfredini et al. [7] and Secchi et al. [8]).

3 Estimating TIM Density Profiles

To characterize the dynamic of the presences in the city by defining daily density profiles (DDPs) is an object of interest. With a DDP we consider the curve representing the number of people in a defined rectangular region at different time periods during the day. We develop a procedure to group similar days in terms of the dynamic of the presences of TIM users considering both the spatial structure and the evolution over the time. By letting X_{it} be the matrix containing the values of the grid at the quarter t of the day i :

- In the first step, using Histogram of Oriented Gradients (Dalal et al. [9], Tomasi [10]), for each i and for each t we extract the vector of features from Z_{it} (a standardized version of the matrix X_{it} such that min/max is in the interval $[0,100]$ for all i and for all t);
- in the second step, we stack together the vectors of features at a daily basis and we perform a traditional k-means cluster analysis where the objects to be clustered are the days and the variables are all the stacked vector of features. With this step we aim at group days with similarities in the spatial structure of the grid;
- in the third step, for each cluster, we perform a functional data analysis (FDA) clustering technique (Bouveyron and Come[11]) to further group days. In this step we group days that are similar in terms of the shape of the curve;
- in the fourth step we define, for each final group, confidence intervals for the DDP, by using functional box plots (Sun and Genton [12, 13]).

Much details can be found in Metulini and Carpita [14]. The output of the procedure is a functional box plot on the DDPs of a group of similar days. For example, by applying the procedure to the rectangular grid defined by latitude 45.516° N - 46.564° N and longitude 10.18° N - 10.245° N (roughly corresponding to the municipality

¹ Works following the single person over the time exist. Unfortunately, those kind of data are available subject to the administration of a survey, that is usually limited in terms of time and usually cover a small portion of the population (Zhaedi and Shafahi [4]).

of Brescia) and composed by 39 x 39 cells, we find that most of the week days of the Summer 2016 belongs to the same group. Functional box plots of that days (Figure 1) highlight the amount of TIM users along different quarters, that varies, by month and by quarter, from a minimum of 30 to a maximum of about 55 thousands of TIM users. These values give us a series of information, for example that the number

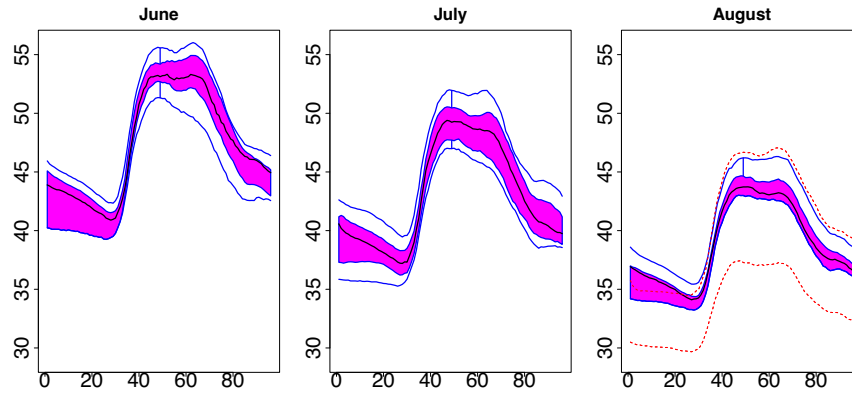


Fig. 1 Functional box plots of the Daily Density Profiles of Summer 2016 (y-axis, in thousands of people), divided by month. Quarters are in the x-axis. Median (black curve), envelope (purple area), 1.5 * envelope (blue curves), outliers (red curves). Our Elaborations.

of people in the city increases during the morning and the afternoon hours and decreases during the night. However, the same values give us no information on the number of total people.

4 The linkage strategy

With the procedure in section 3 we are able to quantify the number of TIM users in the city in selected days and in particular hours of the day. However, to have an idea of the total number of people we have to consider users of other mobile phone companies as well. This data is often unavailable, unless to an onerous cost. An alternative approach is to apply the TIM mobile phone market share coefficient to the number of TIM users to retrieve the total number of people. A country-level estimate being available through “Il Sole 24 Ore” newspaper. This value stands to 30.2 % (2016, December). However, we have reasons to think that TIM market share varies along cities since socio-economic and demographic variables (Table 1) highlight significant differences. To estimate the market coefficient for the municipality of Brescia we employ the following strategy. We compare the data on the number of residents from administrative archives with the number of TIM users on a selected region during a specific hour of the day.

Quantity	Municipality of Brescia	Italy
Per-capita revenues (Euro / year) ¹	23,418	19,514
% foreigners ²	18.5	8.5
Average number of people per family ²	2.11	2.33
Average age ²	45.8	44.7

¹ MEF -Dipartimento delle Finanze (2016)

² ISTAT (2017)

Table 1 Socio-economic and demographic comparison between Municipality of Brescia and Italy.

To the sake of comparison, we assume that, during late evening hours, residential areas are populated just by residents. ISTAT (<https://www.istat.it/it/archivio/104317>) published geographical data called “Basi territoriali e variabili censuarie” in the form of a shape file with data (the so-called `SpatialPolygonDataFrame` in R language). For the municipalities with more than 20,000 residents, ISTAT aggregates the region at a “Sezioni di censimento” (SC) level. To have an idea, the municipality of Brescia has 1,836 SCs. The shape file contains the information on the number of residents by SC. We compute the number of TIM users in each SC by putting together the grid data of the TIM users with the shape files containing the value of residents. The grid cells have regular size while SCs are irregular size polygons. So, to count the number of TIM users in each polygon we apply a weighted scheme based on the portion of the polygon contained in the cell². For example, SC 110, located at latitude 45.544° N and longitude 10.217° N (Figure 2) overlaps with 4 cells. Cell 1, at 9pm - October 28th 2015, counts for 682 TIM users, cell 2 counts for 555 users, cells 3 for 677 and cell 4 751. Moreover, 8.3% of the polygon lies in cell 1, 27.0% in cell 2, 26.4% in cell 3 and 38.2% in cell 4. The number of TIM users in SC 110 is estimated as:

$$TIM\ users = (682 * 0.083 + 555 * 0.270 + 677 * 0.264 + 751 * 0.382) * \frac{area(SC)}{area(cell)}$$

After having computed this number for every SC, we then compute the ratio among TIM users and residents³:

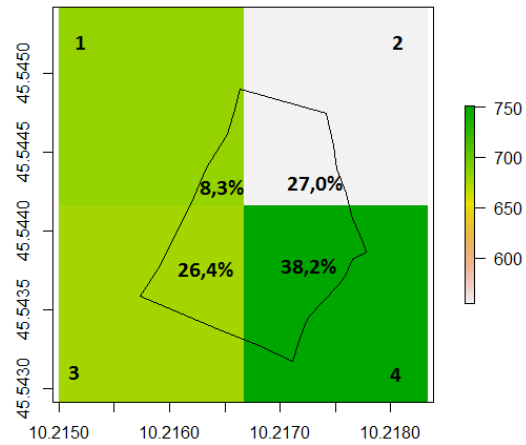
$$Ratio = \frac{TIM\ users}{residents}$$

Some descriptive statistics on the ratio index by SC, evaluated on TIM users at 9pm - October 28th 2015, are reported in Table 2. The median value is consistent with the TIM market share at a country level. However, the values in the right tail of the distribution are extremely large. For example, the 95th percentile stands to 5.567, meaning that the number of TIM users retrieved in that polygon are five times as large as the number of residents. The map in Figure 3 confirms this evi-

² In doing so, we had to convert the coordinate reference system of the spatial polygon object from “UTM” to “longlat”. ISTAT provides shape files in WGS84 international coordinate reference system.

³ In the count of “Residents” we exclude elderly people (>80 years) and children (< 11 years), assuming that we want to consider just those with a smartphone.

Fig. 2 Weighting scheme to assign the number of TIM users to SC 110, located at latitude 45.544° N and longitude 10.217° N.



min	5th percentile	25th percentile	median	75th percentile	95th percentile	max
0.006	0.070	0.139	0.245	0.547	5.567	347.024

Table 2 Quartiles distribution of the *ratio* index in the municipality of Brescia, by SC. 9pm - October 28th 2015.

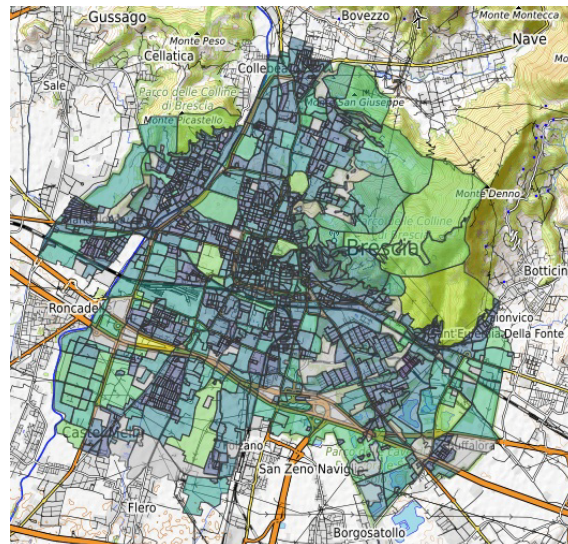


Fig. 3 Map of the *ratio* by SC in the municipality of Brescia, 9pm - October 28th 2015. Colors go from blue (small ratios) to yellow (large ratio).

dence⁴. We find that the ratio value in residential areas is quite homogeneous. Unfortunately, we also notice how TIM assignment of the users in space is affected by the localization of Antennas. So, in many areas (particularly those larger ones where the number of residents is small) the *ratio* index is overestimated. This fact suggests that the Antenna is often localized in large SCs where few people reside. So, people on residential areas are likely tracked where the antenna is (i.e. up to few hundred meters far from where the person is). For this reason, we decide to visually detect in the map residential areas that are located close to anomalies (i.e. where the antenna is). The strategy is to consider the *ratio* index for a selected residential area plus a portion of region that surrounds the area. Residential areas are chosen according to DUSAF (Destinazione d'Uso dei Suoli Agricoli e Forestali) maps (<https://www.dati.lombardia.it/Territorio/Dusaf-5-0-Uso-del-suolo-2015/iq6r-u7y2>). A zoom of the *ratio* map is reported in Figure 4 (Villaggio Sereno residential areas plus a shopping center) and Figure 5 (San Polo residential area plus a factory). The *ratio* index in the first area stands to 0.265 and the same value stands to 0.310 in the area of San Polo.



Fig. 4 Zoom of the *ratio* map for “Villaggio Sereno” area. Latitude 45.505° N - 45.523° N and longitude 10.179° N - 10.194° N.



Fig. 5 Zoom of the *ratio* map for “San Polo” area. Latitude 45.511° N - 45.519° N and longitude 10.235° N - 10.250° N.

5 Conclusions

Mobile phone data can be used to count the number of people in the city considering the dynamics over the time, differently to administrative archives. However, generally, just data on a portion of phone users is available (i.e. just the users of

⁴ A dynamic map can be found at authors personal websites.

a company). For this reason we have developed a method to estimate the market share coefficient of the company at a municipality level - whereas the market share is available only at country level - by using administrative data on the number of residents by “Sezione di censimento”. In doing so we take into account for the bias caused by the location of the antennas. Results are consistent with the market share at a national level.

Acknowledgements Authors are grateful with the Statistical Office of the Municipality of Brescia, with a special mention to Dr. Marco Palamenghi, who kindly supported us with providing the data.

References

1. Zanini, P., Shen, H., & Truong, Y. Understanding resident mobility in Milan through independent component analysis of Telecom Italia mobile usage data. *The Annals of Applied Statistics*, vol. 10(2), pp. 812-833 (2016).
2. Blum, O., Calvo, R., Geospatial data collection and analysis as crucial processes in an integrated census. Israel Central Bureau of Statistics (2001)
3. Xu, S., Flexner, S., Carvalho, V. Geocoding Billions of Addresses: Toward a Spatial Record Linkage System with Big Data (2012).
4. Zahedi, S., & Shafahi, Y. Estimating activity patterns using spatio-temporal data of cell phone networks. *International Journal of Urban Sciences*, 22(2), 162-179 (2018).
5. Carpita, M., & Simonetto, A. Big Data to Monitor Big Social Events: Analysing the mobile phone signals in the Brescia Smart City. *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation*, vol. 5(1), pp. 31-41. (2014).
6. Metulini, R., Carpita, M., On Clustering Daily Mobile Phone Density Profiles, Workshop “High Dimensional Small Data” (Ca Foscari - Venice) (2018).
7. Manfredini, F., Pucci, P., Secchi, P., Tagliolato, P., Vantini, S., & Vitelli, V. Treelet decomposition of mobile phone data for deriving city usage and mobility pattern in the Milan urban region. In *Advances in complex data modeling and computational methods in statistics* (pp. 133-147). Springer, Cham. (2015).
8. Secchi, P., Vantini, S., & Zanini, P. Analysis of Mobile Phone Data for Deriving City Mobility Patterns. In *Electric Vehicle Sharing Services for Smarter Cities* (pp. 37-58). Springer, Cham (2017).
9. Dalal, N., & Triggs, B. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR’05)* (Vol. 1, pp. 886-893). IEEE Computer Society. (2005).
10. Tomasi, C. Histograms of oriented gradients. *Computer Vision Sampler*, pp. 1-6. (2012).
11. Bouveyron, C., Cme, E., & Jacques, J. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, vol. 9(4), pp. 1726-1760. (2015).
12. Sun, Y., & Genton, M. G. Functional boxplots. *Journal of Computational and Graphical Statistics*, vol. 20(2), pp. 316-334. (2011).
13. Sun, Y., & Genton, M. G. Adjusted functional boxplots for spatiotemporal data visualization and outlier detection. *Environmetrics*, vol. 23(1), pp. 54-64. (2012).
14. Metulini, R., & Carpita, M., The HOG-FDA Approach with Mobile Phone Data to Modeling the Dynamic of People’s Presences in the City, *IES 2019 - Statistical evaluation systems at 360°: techniques, technologies and new frontiers* (2019).

Risk-based analyses for non-proportional reinsurance pricing

Analisi Risk-based per il pricing nella riassicurazione di trattati non proporzionali

Fabio Moraldi and Nino Savelli

Abstract The present article introduces a high-level methodology for the reinsurance pricing of non-proportional treaties and applies that to a series of case studies. Such methodology encompasses volatility and capital, and shows the impact that the business uncertainty of primary insurers may have on the prices they have to face for protecting their business through reinsurance.

Abstract *Questo articolo introduce una metodologia per il pricing di trattati non proporzionali nella riassicurazione. Questa metodologia include misure di volatilità e di capitale e mostra come l'incertezza del business delle compagnie di assicurazione impatta sui prezzi che devono sostenere per proteggere il loro business attraverso la riassicurazione*

Key words: Reinsurance, Volatility, Non-proportional treatis

1 General formula for Technical Pricing

introduces the reinsurance technical price C , which is defined as follows:

$$C_i = E(\tilde{X}_i) + \alpha \cdot \sigma(\tilde{X}_i) + \beta \cdot [Y_{i,0.995} - E(\tilde{Y}_i)] + Exp \quad (1)$$

In the following, the meaning and rationale of the different items of the equations are described:

Fabio Moraldi

Assicurazioni Generali S.p.A., Milan

The opinions expressed in this article are those of the author and do not necessarily reflect the views of Assicurazioni Generali S.p.A.

Nino Savelli

Department of applied Mathematics, Finance and Econometrics, Università Cattolica del Sacro Cuore, Milan

- the marker $\tilde{\cdot}$ defines a random variable; claims cost or their functions (e.g. capital) are considered random variables, while items like expenses are assumed deterministic throughout the article;
- i represents the reinsured business; this can refer to a single line of business or, in case of cross-LoBs treaties, multiples lines of business;
- \tilde{X} is the claims cost for the reinsurer; specifically, this refers to the reinsured portion of claims cost and not to the original ground up exposure of the line(s) of business;
- $E(\tilde{X}_i)$ describes the expected value of the reinsurance transaction, i.e. the cost that the reinsurer is expected to pay to the cedant in a given year for the agreed reinsurance treaty;
- $\sigma(\tilde{X}_i)$ describes the standard deviation of the reinsured business; a portion α of this uncertainty is introduced in the technical pricing in order to adapt the reinsurance cost to the volatility of the i -th business reinsured;
- the function $\tilde{Y}_i = f(\tilde{X}_i)$ indicates the capital that the reinsurer has to hold in order to reinsure the business i ; a portion β of this capital is introduced in the technical pricing in order to adapt the reinsurance cost to the additional funding that the reinsurer needs; the degree of dependency between \tilde{X} and \tilde{Y} is a key item that may influence the pricing significantly;
- Exp represents the amount of brokerage and other expenses, including taxes.

Out of the different components of the pricing formula described above, two specific statistical aspects require immediate highlighting.

Firstly \tilde{X}_i , which is, on the one hand side, a function of the cost of ground up claims, in turn depending on their frequency and severity, and, on the other side, a function of the priority and limit of the non-proportional reinsurance scheme. In the simplified case of a single line of business i and of an excess of loss reinsurance with unlimited reinstatements, \tilde{X}_i can be expressed as:

$$\tilde{X}_i = \sum_{j=1}^{\tilde{N}_i} \min[\max(\tilde{Z}_{j,i} - P, 0), L] \quad (2)$$

where:

- \tilde{N}_i is a discrete variable modelling the number of j claims occurring on the line of business i ; throughout the article, we will largely adopt the Poisson model for the claims count; indeed, non-proportional reinsurance is widely used for protecting cedants against large losses, and in turn their frequency is demonstrated by [1] to converge to this model;
- $\tilde{Z}_{j,i}$ is a continuous variable modelling the cost of each individual claim j ; throughout the article, we will adopt a few different models, with predominant use of the Pareto model, thanks to its proven suitability within the reinsurance framework (see also [11] and [7]);
- P is the priority (i.e. attachment point) of the excess of loss reinsurance protecting line i ;

- L is the individual limit (i.e. exhaustion point) of the excess of loss reinsurance protecting line i .

The second item requiring immediate highlighting is the function $f(\tilde{X}_i)$ describing \tilde{Y}_i . Indeed, this is subject to the statistical complexity of \tilde{X}_i , anticipated within the previous paragraph. This implies that such function can be expressed by a closed formula only in a limited number of cases, as recalled within the remainder of the present abstract.

After introducing this pricing definition, section 1 provides examples of application of this approach on two lines of business of a real insurance company. This is pursued by providing means of estimation for $E(\tilde{X}_i)$ and $\sigma(\tilde{X}_i)$ and by assuming that $\tilde{Y}_i = \tilde{X}_i$.

Rationale and sensitivities for the determination of α and β are provided as well. Specifically, it is described that α should be considered as a fixed charge to price the uncertainty, which is solely dependant on the risk appetite of the reinsurer, while β should be considered as a variable charge depending on the cost of the additional capital the reinsurer has to hold when assuming the business i .

2 Implications for multiple lines of business

Extends the framework provided in section 1 by combining the two lines of business in a single reinsurance treaty.

This has the direct effect to introduce the concept of diversification between lines of business in the reinsurance pricing, from the perspective of both the cedant and the reinsurer. In particular, starting from the data of the lines of business presented in section 1, it is shown how material the diversification benefit can be and its impact on the expected value, the standard deviation and the capital. Specifically, it is stressed that, while expected values are always considered as additive, this property is rather lost when considering non-proportional reinsurance treaties, as the charges to certain reinsurance layers depend on the joint behaviour of the losses at high quantiles, this in turn affected by their level of dependency.

Moreover, the aforementioned analysis allows to introduce the concept of asymmetry of information between the cedant and the reinsurer, that by pooling risks end up in different trade-off between reinsurance cost (reinsurance profit) and saved capital (assumed capital).

3 Deep-dive on capital-related implications

further extends the framework provided in sections 1 and 2 by introducing a differentiation between variables \tilde{Y}_i and \tilde{X}_i (i.e. $\tilde{Y}_i = f(\tilde{X}_i)$ where $f()$ is not the identity function).

Specifically, it is assumed that, rather than solely reinsuring the two lines of business from the cedant as in section 2, the reinsurer also underwrites other business from other primary insurers. As a consequence, the reinsurer approaches business i by having a pre-determined level of diversification within its book.

In order to illustrate the concept, one can refer for simplicity to the standard framework provided by Solvency II (i.e. Standard Formula).

Indeed, under this regulatory scheme, and ignoring for simplicity all risks other than premium, the following holds¹:

$$[Y_{0.995} - E(\tilde{Y})] = 3 \cdot \sigma \cdot V \quad (3)$$

where V is the premium-based volume measure and σ is derived according to two consecutive steps, where:

1. $\sigma_i = \frac{\sqrt{(\sigma_{i,prem} \cdot V_i)^2}}{V_i}$ defines the volatility at i line of business level, $\sigma_{i,prem}$ being provided by the regulatory authority;
2. $\sigma = \frac{1}{V} \cdot \sqrt{\sum_{i,l} Corr_{i,l} \cdot \sigma_i \cdot \sigma_l \cdot V_i \cdot V_l}$ defines the volatility at total entity level, obtained by allowing a diversification benefit across i and l lines of business, using the correlation matrix $Corr_{i,l}$ provided by the regulatory authority.

In table 1 we then show an illustrative example to elaborate the aforementioned concept. Specifically, we apply the general formula 1 for a motor vehicle liability (MVL) excess of loss reinsurance with unlimited reinstatements, under the following conditions:

- for case A, the portfolio of the cedant and of the reinsurer is the same and purely composed of risks coming from the motor vehicle liability (MVL) line of business of the cedant;
- for case B, the portfolio of the reinsurer also contains fire risks coming from a different cedant;
- the reinsurer is required to assess its additional risk capital after the MVL reinsurance arrangement on the basis of Standard Formula framework.

The table shows that, while elements like expected loss \tilde{X} , expenses and capital loading β are unchanged in the two scenarios, the capital component $Y_{0.995} - E(\tilde{Y})$ varies in the two scenarios. Indeed, due to the diversification existing in the reinsurer's portfolio in case B, dependent on risks accepted from other entities, the additional capital required by the reinsurer to accept the motor vehicle risks is lower

¹ As the example is based on the Standard Formula framework, the capital is expressed by means of a closed formula which is a function of σ , that in turn is already accounted for in the technical pricing formula (see 1). However, it has to be noted that in the real practice, i.e. when adopting an extended ORSA framework or especially when using an Internal Model, the capital is obtained in a more sophisticated way and it is not a direct function of σ .

Similarly, the Standard Formula framework does not consider any allowance for expected profits within the capital quantification, which can be rather accounted for in an Internal Model framework.

Further insights on these items are provided throughout the article.

than in case A, where cedant and reinsurer exchange exposures on a single given portfolio.

Pricing component	Pricing sub-component	Case A	Case B
$E(\tilde{X}_{MVL})$		5	5
Exp		1	1
α		0%	0%
β		10%	10%
$Y_{MVL,0.995} - E(\tilde{Y}_{MVL})$		30	20.3
	$Y_{MVL,0.995} - E(\tilde{Y}_{MVL})$	30	30
	$Y_{Fire,0.995} - E(\tilde{Y}_{Fire})$	0	19.2
	V_{MVL}	100	100
	V_{Fire}	0	80
	σ_{MVL}	10%	10%
	σ_{Fire}	not relevant	8%
	LoB Correlation	not relevant	25%
Reinsurance Pricing for MVL		8.00	7.03

Table 1 Illustrative example of capital diversification impact in reinsurance pricing

The concepts above imply that the asymmetry of information between the cedant and the reinsurer further increases compared to the analyses described for section 2, and poses a natural question on how the size and geographical diversification of the reinsurer is affecting the offered price.

Within this context, section 3 provides further analyses and sensitivities on β , in order to show the connection between the additional capital required to reinsure the business i and the cost of this capital, this depending on the financial strenght of the reinsurer. Largely capitalized reinsurers would tend to have higher values of β and lower values of $[Y_{i,0.995} - E(\tilde{Y}_i)]$, and viceversa.

4 Conclusions

Summarizes the main conclusions, along with recalling the limitations of the analysis.

Firstly, the relevance of a sound technical pricing is stressed as key for a proper business management on the reinsurer's side, although marking that similar considerations carried out on the cedant's side can also be useful to have an initial understanding on which risks to reinsure and how.

Secondly, a summary and outlook are provided on the effects of diversification and cost of capital in the reinsurance pricing.

With respect to the limitations, the following items are highlighted:

- although being a fundamental instrument, in the real world technical pricing has to be deeply cross-checked with market appetite and conditions; at the same time, it is highly important that a technical pricing culture is widespread in the market to ensure a worldwide sustainable business model;
- the paper does not consider elements like budget or capital restrictions on both cedant's and reinsurer's sides, which also implies the lack of minimum or maximum pricing acceptance thresholds;
- capital analyses are conducted in the perspective of pure underwriting risk, ignoring the effect of other risks, driven by the following considerations:
 - market risk is not considered relevant as, except in extraordinary cases, the acceptance of new reinsurance treaties would not cause the revision of the asset mix of the reinsurer's portfolio;
 - counterparty default risk is not considered relevant as the capital is assessed from the perspective of the reinsurer, which is an ultimate risk taker; specifically, further possible retrocessions performed by the reinsurer would be ad-hoc financial transactions, typically collateralized and therefore not bearing counterparty default risk;
 - operational risk is not considered due to materiality;
 - reserving risk is not considered to the selected mono-annual view of the risk capital.

References

1. Embrechts, P., Kluppelberg, C., Mikosch, E.: Modelling extremal events: for insurance and finance. Springer (1997)
2. Di Gropello, G., Gionta, G.: Manuale di riassicurazione. LINT Editoriale (2004)
3. Daykin, C.D., Pentikainen, T., Pesonen, M.: Practical risk theory for actuaries. Chapman & Hall (1994)
4. Savelli, N.: Solvency and traditional reinsurance for non-life insurance. 6th International Congress on IME (Insurance: Mathematics and Economics) (Lisbon) (2002)
5. Clemente, G.P., Savelli, N., Zappa, D.: The impact of reinsurance strategies on capital requirements for premium risk in insurance. *Risks* (2015)
6. Schmutz, M., Doerr, R.: The Pareto model in property reinsurance: formulas and applications. Swiss Re (1998)
7. Schmitter, H., Butikofer, P.: Estimating property excess of loss risk premiums by means of the Pareto model. Swiss Re (1997)
8. Gussisberg, D.: Exposure rating. Swiss Re (2004)
9. Chenut, X.: Property XL rating: a reinsurance pricing tool combining experience and exposure rating, 38th ASTIN Colloquium (Manchester) (2008)
10. Clark, D.R.: Basics of reinsurance pricing. CAS Study Note (1996)
11. Rytgaard, M.: Estimation in the Pareto distribution. *ASTIN Bulletin* 20 (1990)
12. Klugman, S., Willmot, G., Panjer, H.: Loss models: from data to decisions. Wiley (2008)

A Simplified Efficient and Direct Unequal Probability Resampling

Un semplice Ricampionamento, efficiente e diretto per campioni a probabilità variabili

Federica Nicolussi, Fulvia Mecatti and Pier Luigi Conti

Abstract In this paper, a new resampling technique for sampling designs with unequal inclusion probabilities is proposed. The basic idea is to use a resampling design based on *ppswor*. Its main properties are studied, and its relationships with other resampling methodologies are discussed.

Abstract In questo lavoro si introduce una tecnica di ricampionamento valida per disegni campionari con differenti probabilità di inclusione. L'idea di base è di usare un disegno di ricampionamento di tipo *ppswor*. Le principali proprietà del metodo sono studiate, e le relazioni con altre metodologie di ricampionamento sono discusse.

Key words: resampling, finite population, sampling design, *ppswor*.

1 Background and Contribution

Resampling algorithms are simple and general tools for assessing estimators' accuracy via variance estimation and for producing confidence intervals and p-values. Resampling provides numerical solutions in non-standard challenging inferential setups so that it has a special appeal for dealing with complex sampling designs for finite population. These include the popular without replacement probability proportional to size (π psWOR) sampling, where every population unit is assigned a specific probability to be included into the final sample, defined as proportional to an available (positive) covariate with the role of auxiliary variable. The Bootstrap,

Federica Nicolussi
Università degli Studi di Milano, e-mail: federica.nicolussi@unimi.it

Fulvia Mecatti
Università degli Studi di Milano-Bicocca, e-mail: fulvia.mecatti@unimib.it

Pier Luigi Conti
Sapienza Università di Roma e-mail: pierluigi.conti@uniroma1.it

likely the most used resampling method originally proposed by Efron [8] for *iid* sample data, does not work in sampling from finite populations, since it cannot deal with the dependence among sample units due to the sampling design. Several modified techniques have been proposed to overcome this problem. In a recent extensive review [12] such proposals are classified into three groups

1. methods based on a *pseudo-population*, where sample units are first used to construct a *replicate* of the parent population and then bootstrap samples are selected into the resulting pseudo-population. Main proposals in this class are [9], [5], [4], [11] and more recently [6] and [7];
2. *direct* bootstrap methods where bootstrap samples are directly selected from the (original) sample or a *re-scaled* version of it. Main proposals in this class are [14], [18] and recently [1];
3. *weighted* bootstrap methods, where a new set of weights is provided to produce bootstrap estimates, by adjusting the (original) design weights. Main contributions of this third type are [15] and [2].

In a recent paper Conti et al. [7] provide a general theory for finite population resampling based on pseudo-population by also proving its asymptotic correctness. The main contribution parallels the asymptotic justification by Bickel and Friedman [3] for the classical *iid* Efron bootstrap.

In this paper a new π psWOR resampling is introduced which is a simplified and computationally more efficient version of the asymptotically correct bootstrap by Conti et al. [7]. The new proposal presents several advantages w.r.t. the large available literature on the topic. First of all, it represents a unified approach to re-sampling complex samples from finite population. It is in fact a method based on a *pseudo-population*, asymptotically correct according to Conti et al. [7]. However, at the same time it is both a direct bootstrap and a weighted bootstrap, for allowing to select bootstrap samples directly from the original sample on the basis of an appropriate (bootstrap) weighting system. Secondly, it is computationally efficient because it does not require the actual construction of a pseudo-population. In the third place, it is important to notice that the real application of a finite population re-sampling usually (and certainly for existing methods included in group 1.) involves some sort of rounding or re-scaling, either randomized or systematic, which would affect the entire bootstrap performance and ultimately the expected properties of the released bootstrap estimates. The resampling we are proposing does not need any arbitrary rounding and it admits underlying pseudo-population of any size possibly non-integer, along with any real value for the bootstrap weights. A greater precision and possibly efficiency gains are expected as a consequence. Finally, our resampling is very simple to implement, since it requires a unique basic re-sampling design whatever π psWOR design had generated the available to-be-bootstrapped sample.

2 Notation and Preliminaries

Let \mathcal{U}_N be a finite population of unit $i = 1 \dots N$ from which a sample s is selected under a given design and with pre-fixed size n . Let D_i be the sample membership indicator, i.e. a random variable taking value 1 if $i \in s$ and 0 otherwise, with $n = D_1 + \dots + D_N$. The (design) expectation $\pi_i = E[D_i] = P(i \in s)$ is the first order inclusion probability. Let \mathcal{Y} be the study variable and \mathcal{X} be an available positive auxiliary variable, with y_i and x_i their value for each population unit, $t_y = \sum_i y_i$ and $t_x = \sum_{i=1}^N x_i$ their population totals. A π psWOR sampling design is known to be highly efficient whenever \mathcal{Y} is expected to be in a relation of approximate proportionality with \mathcal{X} , so that π_i are set to be proportional to the auxiliary variable $\pi_i = nx_i/t_x$, $i = 1 \dots N$. Let $\theta = \theta(F_N)$ be the population quantity to be estimated, where F_N denotes the population distribution function of \mathcal{Y} . We focus on the familiar and often used class of estimators that are expressed as functional of an estimator of F_N , namely $\hat{\theta} = \theta(\hat{F})$. Such class includes both the popular Horvitz-Thompson and Hájek estimators, respectively given by the following choices to estimate F_N

$$\hat{F}_{HT}(y) = \frac{1}{N} \sum_{i=1}^N \frac{1}{\pi_i} D_i I_{(y_i \leq y)}, \quad \hat{F}_H(y) = \frac{\sum_{i=1}^N \frac{1}{\pi_i} D_i I_{(y_i \leq y)}}{\sum_{i=1}^N \frac{1}{\pi_i} D_i}. \quad (1)$$

Whatever complex both the sampling design and the analytical structure of θ , a resampling algorithm would produce a (Monte Carlo) estimate of the variance of $\hat{\theta}$ as well as confidence intervals for θ . In Conti et al [7] conditions are given for a resampling algorithm to be asymptotically correct, which implies to be based on a pseudo population and to comprise the following basic steps, where the familiar $star^*$ notation is adopted to denote bootstrap quantities:

0. Construct a pseudo-population $\mathcal{U}_{N^*}^*$ by replicating (a chosen number) N_i^* of times the values $\{y_i, x_i\}$ associated with every sampled unit $i \in s$. From now on y_k^* and x_k^* will indicate the study and auxiliary values included into the pseudo-population of size N^* , such that N_i^* units $k \in \mathcal{U}_{N^*}^*$ would be of Type i , with $i \in s$ and $N^* = \sum_{i \in s} N_i^*$. Finally define the pseudo-population distribution function as

$$F_{N^*}^*(y) = \frac{1}{N^*} \sum_{k=1}^{N^*} I_{(y_k^* \leq y)} = \sum_{i=1}^N \frac{N_i^*}{N^*} D_i I_{(y_i \leq y)}, \quad y \in \mathbb{R}. \quad (2)$$

1. Generate M independent bootstrap samples s^* of size n (M chosen sufficiently large) by selecting from $\mathcal{U}_{N^*}^*$ under a (re)sampling design guaranteeing first order inclusion probabilities $\pi_k^* = nx_k^*/t_x^*$ where $t_x^* = \sum_{k=1}^{N^*} x_k^* = \sum_{i \in s} N_i^* x_i$.
2. For each bootstrap sample s_m^* compute $\hat{F}_{H,m}^*$ according to the right term in (1) and thus compute the replicate $\hat{\theta}_m^* = \theta(\hat{F}_{H,m}^*)$, $m = 1 \dots M$.
3. Compute the M quantities

$$Z_{n,m}^* = \sqrt{n}(\hat{\theta}_m^* - \theta^*) = \sqrt{n}(\theta(\hat{F}_{H,m}^*) - \theta(F_{N^*}^*)) \quad m = 1, \dots, M. \quad (3)$$

Note that (3) provides a bootstrap distribution of $\hat{\theta} = \theta(\hat{F})$ - simulated upon M runs - with empirical distribution function given by $\hat{R}_{n,M}^*(z) = \frac{1}{M} \sum_{m=1}^M I_{(Z_{n,m}^* \leq z)}$, $z \in \mathbb{R}$ and corresponding p th quantile defined as

$$\hat{R}_{n,M}^{*-1}(p) = \inf\{z : \hat{R}_{n,M}^*(z) \geq p\}, \quad 0 < p < 1. \quad (4)$$

4. Compute the variance of (3)

$$\hat{S}^{2*} = \frac{1}{M-1} \sum_{m=1}^M (Z_{n,m}^* - \bar{Z}_M^*)^2 = \frac{n}{M-1} \sum_{m=1}^M (\hat{\theta}_m^* - \bar{\theta}_M^*)^2 \quad (5)$$

where $\bar{Z}_M^* = \frac{1}{M} \sum_{m=1}^M Z_{n,m}^*$ and $\bar{\theta}_M^* = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m^*$, which is a bootstrap (point) estimate of the variance of estimator $\hat{\theta}$.

6. Finally bootstrap confidence intervals (CI) can be computed; for instance based on percentiles (4)

$$\left[\hat{\theta} - n^{-1/2} \hat{R}_{n,M}^{*-1}(1 - \alpha/2), \hat{\theta} - n^{-1/2} \hat{R}_{n,M}^{*-1}(\alpha/2) \right] \quad (6)$$

and based on Standard Normal percentiles and the bootstrap variance estimate at step 3.

$$\left[\hat{\theta} - n^{-1/2} z_{\alpha/2} \hat{S}^*, \hat{\theta} + n^{-1/2} z_{\alpha/2} \hat{S}^* \right]. \quad (7)$$

3 The Proposed Method

With the purpose of contributing a new π psWOR resampling which is a simplified and computationally more efficient version of the asymptotically correct bootstrap recalled above, we aim at compressing the initial steps 0. and 1. in a unique simplified and significantly less time-consuming step. Toward this goal a promising starting point has been a recent Quatember proposal [13]. It focused on a pseudo-population based on the quite natural choice

$$N_i^* = \pi_i^{-1} = t_X / (n x_i) \quad i \in s \quad (8)$$

which are usually non-integer numbers so that some sort of rounding is needed, the most popular being the Holmberg randomization [11]

$$N_{i,Holm}^* = \lfloor \pi_i^{-1} \rfloor + \text{Bern}(\pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor). \quad (9)$$

The ingenious Quatember solution has two main properties: first of all, the actual construction of the pseudo-population is unnecessary and in fact skipped, so that N^* and every N_i^* are allowed any real number non-necessary integer. Important simplifications realize as a consequence: both the construction of the pseudo-

population and any rounding would be avoided, whether randomized or otherwise. Second the M bootstrap samples s_m^* are selected directly from the (original) sample s by a simple draw-by-draw (with replacement, WR) design, related to the so called drawing-probability-proportional to size (pps). Quatember re-sampling design has initial probability of selecting a unit of Type i , either from the underlying $\mathcal{U}_{N^*}^*$ or directly from s , set equal to

$$p_{i,Q}^* = x_i/t_X^*. \quad (10)$$

However it can be proved that Quatember method is not asymptotically correct, according to Conti et al. [7]. This is essentially because it fails to address the requirement for the re-sampling inclusion probabilities to be proportional to the auxiliary variable, precisely $\pi_k^* = nx_i/t_X^*$ for all N_i^* units $k \in \mathcal{U}_{N^*}^*$ of Type $i \in s$.

By benefit upon the main idea in Quatember first proposal and by retaining its simplicity and computational efficiency, we now propose some modified versions of $p_{i,Q}^*$ able to address, at least approximatively, such requirement which would lead to asymptotically correct resampling methods beside simplified and efficient. It is important to notice that inclusion probabilities for pps design are not proportional to p_i s, and do not have an expression in closed form (see for instance [10], p. 95). We then rely upon three useful approximations, each relating the initial selection probability p_i to the first order inclusion probability π_i . By conditioning on the re-sampling first order inclusion probability to equate nx_i/t_X^* , we derive

$$p_{i,R1}^* \approx \log \left(1 - \frac{nx_i}{t_X^*} \right) / \sum_{l \in s} N_l^* \log \left(1 - \frac{nx_l}{t_X^*} \right) \quad (11)$$

as a first solution based on [16]. A second solution, computationally heavier, is based on [17] and can be computed via the following iterative algorithm:

0. Set $m = 0$, $\pi_{(i)}^*(m) = \pi_{(i)}^*$, $i \in s$, and take a threshold $\delta > 0$.

1. Compute

$$p_i^*(m) = \log \left(1 - \pi_{(i)}^*(m) \right) / \sum_{l \in s} N_l^* \log \left(1 - \pi_{(l)}^*(m) \right), \quad i \in s.$$

2. Compute $\xi_n^*(m)$ as the solution of the equation:

$$\sum_{i \in s} N_i^* (1 - \exp \{-p_i^*(m)t\}) = n.$$

3. Compute

$$\pi_i^*(m+1) = 1 - \exp \{-\xi_n^*(m)p_i^*(m)\}, \quad i \in s. \quad (12)$$

4. Set $m \rightarrow m+1$. If $|\pi_i^*(m+1) - \pi_i^*| < \delta$ for every $i \in s$, then go to Step 5. Otherwise, go to Step 1.

5. Set

$$p_{i,R2}^* = p_i^*(m), \quad i \in s. \quad (13)$$

A third option, based on [10], has led to

$$p_{i,H}^* = \frac{x_i}{t_X^*} \left\{ 1 + \frac{1}{2} \frac{n-1}{n} \left(\frac{nx_i}{t_X^*} - \bar{\pi}^* \right) \right\} \quad (14)$$

where $\bar{\pi}^* = n^{-1} \sum_{i \in S} N_i^* \pi_i^{*2}$.

Finally we considered a fourth solution based on adjusting the choice (8) rather than the initial selection probability. Notice that eqn. (8) leads to the important property $t_X^* = t_X$, *i.e.* the resulting pseudo-population is calibrated w.r.t. the (real) total of the auxiliary variable. On the other hand, neither (8) nor its randomized version (9) satisfy $\sum_{i \in S} N_i^* = N$, *i.e.* the pseudo-population is not calibrated w.r.t. the population size. Our fourth option is based on fostering a pseudo population calibrated w.r.t. both the (real) population size and (real) total of \mathcal{X} . Such double calibration (DCal) is reached by replacing either equation (8) or (9) by

$$N_i^* = \frac{1}{\pi_i} + \frac{(N - \sum \pi_i^{-1})(\sum x_i^2)}{n(\sum x_i^2) - (\sum x_i)^2} - \frac{(N - \sum \pi_i^{-1})(\sum x_i)}{n(\sum x_i^2) - (\sum x_i)^2} x_i \quad (15)$$

that is the exact solution of a quadratic constrained optimization problem, and can be any real number, whether integer or not.

Our new method consists in replacing both steps 0. and 1. of the asymptotically correct resampling algorithm given in Section 2, by the following unique simplified and computationally more efficient step

1. Generate M independent bootstrap samples s^* of size n (M chosen sufficiently large) by selecting from s under a draw-by-draw WR (re)sampling design with conditional probability of selecting a unit of Type i at the j th bootstrap draw given by

$$p_j(\text{Type } i | s_{j-1}^*) = \frac{\max\{0, (N_i^* - h_{i,j-1})x_i\}}{t_X^* - \sum_{l \in s} h_{l,j-1}x_l} \quad j = 2 \dots n \quad (16)$$

where s_{j-1}^* denotes the bootstrap sub-sample informed by the previous $j-1$ draws, $h_{i,j-1}$ is the number of units of type i selected in the first $j-1$ draw. Notice that, at the first draw ($j=1$) would hold any of the options illustrated in the present section, either equations (11), (13), (14) or (10) joined with (15).

4 Preliminary Empirical Evidence

A preliminary simulation has been carried out with the purpose of empirically testing the performance of our simplified resampling method according to each of the

4 alternative options illustrated in Section 3, and to compare it with some main competitors available in the literature. The simulated scenarios are composed by two populations of increasing size $N = 200, 400$. The study and auxiliary variable \mathcal{Y} and \mathcal{X} were generated according to the same model in [1] leading to circa 80% of correlation. For each scenario, 1000 samples were simulated under a Pareto π ps design with 20% sampling fraction. Focusing of the population mean $N^{-1} t_y$ as the quantity θ to be estimated, we simulated two familiar estimators: the unbiased Horvitz-Thompson estimator $\hat{\theta}_{HT} = N^{-1} \sum_{i \in S} y_i \pi_i^{-1}$ and the more efficient and asymptotically unbiased Hájek estimator $\hat{\theta}_H = (\sum_{i \in S} \pi_i^{-1})^{-1} \sum_{i \in S} y_i \pi_i^{-1}$.

For each simulated sample, $M = 1000$ bootstrap runs were performed under 7 different resampling methods: 4 proposed in this paper plus 3 competitors, as described in Table 1. The first competitor, dubbed *Holm*, consists of the asymptotically correct resampling algorithm recalled in Section 2 with the pseudo-population constructed via the Holmberg randomization. It is thus interesting to compare with the resampling proposed here which aim at an equivalent resampling but computationally more efficient for avoiding both the actual construction of the pseudo-population and any rounding. The second competitor, *DirAT* for short, has been recently proposed in the literature as a direct bootstrap neither based on a pseudo-population nor requiring $\pi_i^* \propto x_i$ for every sample unit $i \in s$. Thus, it appears interesting to compare with our methods w.r.t the statistical properties of the bootstrap estimate provided. Finally, the third competitor briefly indicated by *Q*, is the original proposals by Quatember which has been the starting point for developing our new proposal for a simplified, computationally more efficient and yet asymptotically correct resampling.

Table 1 7 Simulated Resampling Methods

Method	Main features	Reference
<i>Holm</i>	resampling from $\mathcal{U}_{N^*}^*$ with $N_i^* = \lfloor \pi_i^{-1} \rfloor + \text{Bern}(\pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor)$ under the same original sampling design with $\pi_k^* = n x_k^* / t_{X^*}^*$	[11]
<i>DirAT</i>	no $\mathcal{U}_{N^*}^*$, direct resampling into s under a combination of special designs	[1]
<i>Q</i>	direct resampling into s under WR <i>pps</i> design with initial draw probabilities (10) and $N_i^* = \pi_i^{-1}$	[13]
<i>R1</i>	direct resampling into s under WR <i>pps</i> design with initial draw probabilities (11) and $N_i^* = \pi_i^{-1}$	new
<i>R2</i>	direct resampling into s under WR <i>pps</i> design with initial draw probabilities (13) and $N_i^* = \pi_i^{-1}$	new
<i>H</i>	direct resampling into s under WR <i>pps</i> design with initial draw probabilities (14) and $N_i^* = \pi_i^{-1}$	new
<i>DCal</i>	direct resampling into s under WR <i>pps</i> design with initial draw probabilities (10) and N_i^* as in (15)	new

Expected results from our preliminary simulation are

- a significant/dramatic outperformance of any of the new methods $R1, R2, H$ or $Dcal$ over both $Holm$ and $DirAT$ w.r.t. the computational time needed for producing bootstrap estimate;
- an essentially equivalent performance of any new methods $R1, R2, H$ or $Dcal$ as compared to $Holm$ w.r.t. to the properties of the final bootstrap estimates with possibly slight gains due to the possibility to avoid the (randomized) rounding;
- an improvement of all new methods over Q as N and n increases for moderate sampling fraction;
- differences between the performance of the 7 simulated resampling methods able to suggest recommendations for practical application (beside the computational efficiency).

References

1. Antal E., Tillé Y.: A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, **206**, 534–543 (2011)
2. Beaumont J.-F., Patak Z.: On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling. *International Statistical Review*, **80**, 127–148 (2012)
3. Bickel P.J., Freedman D.: Some asymptotic theory for the bootstrap. *The Annals of Statistics*, **9**, 1196–1216 (1981)
4. Booth J.G., Butler R.W., Hall P.: Bootstrap methods for finite populations. *Journal of the American Statistical Association*, **89**, 1282–1289 (1994)
5. Chao M.-T., Lo S.-H.: A bootstrap method for finite population. *Sankhya*, **47**, 399–405 (1982)
6. Chauvet G.: Méthodes de bootstrap en population finie. Ph.D. Dissertation, Laboratoire de statistique d'enquêtes, CREST-ENSAI, Université de Rennes 2 (2007)
7. Conti P.L., Marella D., Mecatti F., Andreis F.: A unified principled framework for resampling based on pseudo-populations: asymptotic theory. Technical Report, Arxiv 1705.03827. Submitted for publication, (2017)
8. Efron B.: Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1–26 (1979)
9. Gross S.T.: Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 181–184 (1980)
10. Hájek J.: *Sampling from a finite population*, Marcel Dekker, New York (1981)
11. Holmberg A.: A bootstrap approach to probability proportional-to-size sampling. *Proceedings of the ASA Section on Survey Research Methods*, 378–383 (1998)
12. Mashreghi Z., Haziza D., Léger C.: A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, **10**, 1–52 (2016)
13. Quatember A.: The Finite Population Bootstrap - from the Maximum Likelihood to the Horvitz-Thompson Approach. *Austrian Journal of Statistics*, **43**, 93–102 (2014)
14. Rao J.N.K., Wu C.F.J.: Resampling inference with complex survey data. *Journal of the American Statistical Association*, **83**, 231–241 (1988)
15. Rao J.N.K., Wu C.F.J., Yue K.: Some recent work on resampling methods for complex surveys. *Survey Methodology*, **18**, 209–217 (1992)
16. Rosén B.: On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, **62**, 159–191 (1997)
17. Rosén B.: On inclusion probabilities for order π_{ps} sampling. *Journal of Statistical Planning and Inference*, **90**, 117–143 (2000)
18. Sitter R.P.: A resampling procedure for complex data. *Journal of the American Statistical Association*, **87**, 755–765 (1992)

Labour Law: Machine vs. Employer Powers *Diritto del lavoro: Macchina vs. Poteri datoriali*

Antonella Occhino – Michele Faioli¹

Abstract - Work entails important theoretical questions if it is linked to the most advanced technology, as broadly diffused in Europe and so called Industry 4.0 and Gig-Economy. Industry 4.0 mainly refers to productions/manufacturing industries, while Gig-Economy to distribution/services industries. Industry 4.0 focuses on labour law as to its internal flexibilities, such as salary, jobs tasks, working hours. On the other hand Gig-Economy focuses on labour law as to its external flexibilities, such as types of work, self-employment vs. employment relations, disguised forms of work, bogus self-employment. Main problems and possible solutions also in terms of labour regulations are examined, including *de jure condendo* legislation and likely unions' strategies, within a legal comparative approach.

Abstract – *Il lavoro pone questioni teoriche importanti in relazione alle tecnologie avanzate, cd. "Industry 4.0" e "Gig-Economy", tenendo presente il quadro europeo. Industry 4.0 è prevalentemente collegato alla manifattura, mentre Gig-Economy è maggiormente collegato alla distribuzione e ai servizi alla persona. Industry 4.0 si innesta nel diritto del lavoro dal punto di vista delle flessibilità interne, anche in una logica di riforma su salario, mansioni, orario. Gig-Economy richiede al diritto del lavoro di ritornare sui temi delle flessibilità esterne che attengono al tipo di lavoro, autonomo o subordinato, e alle relative tutele. Problemi e possibili soluzioni anche in termini di regolazione sono oggetto di esame, compresa la prospettiva della legislazione de jure condendo e delle strategie sindacali, in una logica comparativa.*

Key words: Artificial Intelligence - Robotics – Machine Learning – Labour Protections – Collective Bargaining – Industry 4.0 – Gig-Economy

1 Industry 4.0 and Gig-Economy. Labour Law Approaches

A new business model is being applied to work, partially overtaking the classic, maybe divisive, categories through which the phenomenon is analysed according to the organizational theory (by way of example, organisation – subordination – powers – rights – obligations). Work entails important theoretical questions where linked to the most advanced technology, in particular face to Industry 4.0 and Gig-Economy. Industry 4.0 is broadly related to productions/manufacturing industries, while Gig-Economy is specifically dealt with distribution/services industries. Labour law is needed to intervene after Industry 4.0 more on the internal flexibility regulation: salary, jobs tasks, working hours. Gig-Economy makes labour law more invest on the external flexibility regulation: types of work, self-employment vs. employment relations, disguised forms of work, bogus self-employment.

¹ Antonella Occhino, Full Professor of Labour Law, Università Cattolica del Sacro Cuore - Milan (IT), antonella.occhino@unicatt.it; Michele Faioli, Researcher of Labour Law, University of Roma Tor Vergata – Rome (IT), michele.faioli@uniroma2.it.

2 A Gig-Workers Statute after EU Directive April 16, 2019.

2.1 Workers vs. Employees?

As a first approach to such research themes, we investigate the on-demand platforms and any forms of work performed by gig-workers. Such work could be subject to the idea that platform business aims at avoiding the regulatory obligations which most jurisdictions impose on employers. In other words, a sort of strategic dissimulation seems to be carried out by employers.

Assuming that subordination itself essentially depends on the effective exercise of employment powers, any worker should be protected by labour standards at the only condition to perform under the direction and control of another person, i.e. the employer, as an employee *stricto sensu*. By contrast, autonomous workers traditionally are excluded by the scope of labour legislation, as free lance.

A third way is needed for gig-workers, first to qualify them as workers *tout court*, second (and consequently) to provide them for some rights (and obligations). Eventually, given the main notion dividing between employees and free lance, gig-workers either ask to be protected as employees, or not. Indeed, in order to qualify subordinated workers a set of formal criteria has been provided in any national legal system, but its application is now stressed by the technological modernisation (cf. Mengoni, 1985).

Observing the different realities we face with in such field, we understand that Gig-Economy innovation is also reshaping our economy by incentivising to move away from the hierarchical organisation of business in large corporations, and to embrace open markets, where the real difference between employees (i.e. subordinated workers) and others seems to be ever more uncertain.

Indeed for a long time in several European States the traditional dividing between employees and non-employees has already supposed to be replaced by a new tri-dimensional framework where there are “workers” *tout court*, beside employees (i.e. subordinated workers) and free-lance (i.e. autonomous workers), with workers claiming for the almost same protection as that provided for employees.

The most recent labour law studies have focused on the classification of working activities carried out by the so-called “riders”, engaged through platforms, differentiating their performances between self-employment and employment. In doing that, authors either focus on the re-introduction of other types of working arrangements (so-called “quasi-subordinate” or “coordinated” employment), or directly prefer pointing to the inadequacy of domestic legal frameworks in providing protections for gig-workers (i.e. riders) as to wage, working time, monitoring, union freedoms and strike (as to the differentiation, Prassl, 2017; Davidov, 2016; De Stefano, 2016; Means, Seiner, 2016; Rogers, 2016; Stone, 2016).

2.2 Employers vs. Agency Work?

The French legal approach shows an opting-out of this dividing, having recently introduced by legislation a sort of social responsibility in order to protect intermediate work, such as that of gig-workers. Article L7342-1 (introduced by Article 60 of Law no. 1088 dated 8 August 2016), envisages a specific form of intermediation with social responsibility upon the digital platform: “Lorsque la plateforme détermine les caractéristiques de la prestation de service fournie ou du bien vendu et fixe son prix, elle a, à l’égard des travailleurs concernés, une responsabilité sociale qui s’exerce dans les conditions prévues au présent chapitre”.

The Italian legal framework is still under a complex reforming and reflecting process. A first idea is broadly shared, which encompasses gig-workers, also upon the indication of labour judges (*Corte di Appello di Torino*, Feb. 26, 2019, no. 26), within the personal scope stated by Article 2 of Law no. 81 dated 15 June 2015 (i.e. special regime extending to any self-employment relations, having features of employment relations, with the application to gig-workers of classic labour law protections – wage, social security, maternity, working hours, etc. (Occhino, 2018; Faioli, 2018; Carabelli, Spinelli, 2019).

Under another perspective, jobs carried out through digital platforms – in the specific case of working activities performed for gig economy companies delivering goods (e.g. *Deliveroo*, *Foodora*, *Just Eat*, etc.) or providing services to individuals and households (e.g. *Vicker*, *Task Rabbit*, etc.) – are assumed to be likely to fall under temporary agency work as per the Italian Laws nos. 81 dated 15 June 2015 and 276 dated 10 September 2003 (Treu, Sacchi, 2017; Faioli, 2018).

Eventually whereas the digital platform co-coordinates, co-manages, co-monitors, along with the client (restaurants, pizzeria, etc.) and, in some case, sanctions the worker/rider with a view to meeting a user request (e.g. a restaurant or coffee bar which joins the platform) in relation to the delivery of food to clients, in this way and consequently, restaurant managers do not avail themselves of an employee but of a supposed temporary agency worker by accessing the digital platform.

This may entail a double conceptual shift. On the one hand, *de iure condendo*, if the digital platform became a temporary work agency, it would be subject to the provisions set out in Laws nos. 81/2015 and 276/2003 (with some necessary law amendments concerning sanctions and references to collective bargaining). On the other hand (and this is the most important aspect of the present analysis), such a digital platform would be part of a sort of unitary network of active labour market policies, being enabled to take part in job placement activities and matchmaking (i.e. matching of labour demand and supply) in relation to both traditional jobs (as already known) and Gig-Economy jobs.

2.3 European Protection after EU Directive April 16, 2019.

Leaving aside how the work carried out by gig-workers through digital platforms (in case of riders, specifically aimed at delivering goods) could be qualified, we should highlight that, also *de jure condendo*, the Italian and EU legislator should provide for by gig-workers a core labour standard protection, in order to:

- (i) extend to workers of such digital platforms a set of already established labour law, social security, and union protection measures, with necessary reference to collective bargaining for the definition of some aspects concerning wages and labour costs
- (ii) take the opportunity to improve employability through the promotion of smart and efficient matchmaking mechanisms. Such theory stems from the idea that, as to work performed through digital platforms, it would be useful to introduce a specific set of rules making reference to the legislation on temporary agency work (market design).

In this open context the very recent EU Social Directive April 16, 2019 has provided for by workers (hopefully including gig-workers) the extension of the set of national core labour standards, to be implemented and enforced in each Member State by 2022 with the specific aim at equalising protections for atypical workers to those already provided for by typical ones in each national (internal) legal European orders.

3 Artificial Intelligence in Employment Relationships

3.1 Artificial Intelligence and the Machine

If work has already been considered “incomplete” (non-hyphenated word) because of the theoretical economic incompleteness of the contract of employment - given that work effort bargain and labour capacity cannot easily be specified ex ante - it should be expected to be all the more “in-complete” in the future, due to the peculiar legal relationship which will come to life among (i) the worker, (ii) the employer, and (iii) the machine endowed with artificial intelligence (intelligent machine).

Law and collective bargaining provisions adapt to technology inasmuch as they are shaped by the latter. Though such provisions do not determine a separated system, labour regulation has been adapting a very important part of its logical and legal tools on technology (Faioli, 2018).

In the close future, even the object of the employment contract will be likely to complement with the monitoring and regulatory action carried out by machines on those job tasks, with the aim at impeding workers to be stripped of by the most advanced systems created by micro-electronics and computers and related to automation of industrial processes, logistics, transportation, data elaboration, accounting, quality and cost monitoring, medical interventions, biological and chemical processes, waste disposal, selection of materials, and so on.

This contribute sets out to provide some answers to the following questions.

- How is work in a company changing (or has it already changed) in view of the most recent and innovative technological developments, as well as of (also foreign) investments betting on artificial intelligence?
- How will work labour-regulations change in Europe in the close future, taking into account that automated and smart technology develops at a fast pace in all productive sectors?
- To what extent and how should EU labour-regulations deal with the fact that intelligent machines, endowed with deep learning, will indirectly decide, and monitor, patrol, indicate, define, workers’ job tasks?

3.2 Artificial Intelligence and Big Data

Within such a complex framework, a worker (or group of workers) with specific job tasks and covered by a specific employment contract (including a specific classification system stemming from national collective bargaining) is concerned by a wealth of automated and computerised technology innovations. These impact on his/her as an individual, and require specific knowledge (through something less than a “replacement of workers by machines”) also in terms of coordination with other workers, in the same workplace or somewhere else.

From this perspective, the objective is the search for a methodology aimed at outlining a theory of job tasks in a changing industrial world under pressure from the intelligent machine, which indirectly (throughout algorithms) nearly teaches itself and “coordinates” working activities.

Over time, the picture will become more complex as company-related events will be measured, thus generating huge quantities of archived data (big data), until such data will furthermore be analysed by systems aimed at supporting decision making and based on artificial intelligence technology (so called cognitive systems).

It should be added that there will be the possibility to communicate such data with real-time remote control from one end of the world to the other. As a result, employers’ powers, business networks, and even the notion of

“employer” will be likely to be re-shaped. Connections between the intelligent machine and its big data, on the one hand, and on the other hand employers powers in action should be more precisely investigated.

The idea of the machine as nearly like a “third element” of the contract may help to investigate the legal issues relating to the compatibility between job tasks, staff classification, and the most innovative organisational models (accelution, lean production, team working, etc.) (Faioli, 2018). Eventually these models are included into new organisational processes, which apparently seem to be directly coordinated by machines endowed with artificial intelligence.

4 Technology and Collective Bargaining Functions

4.1 Collective Agreements and Minimum Wage (as to Gig-Economy)

Observing gig-workers and unions activities, there are, at least at the moment, partial outputs for the unions’ objective to set – through collective bargaining – a wage level for gig-workers or other forms of protections, except those expressed by the NCBA (National Collective Bargaining Agreement) for Logistics and Distribution in 2018, which eventually have been extended to riders by a recent Italian labour judges decision (*Corte di Appello di Torino*, Feb. 26, 2019, no. 26) (Occhino, 2019).

The most recent strategic position of main Italian unions deals with the idea to apply protections compared with employees in such industrial sector, such protections being *pro rata* and related to specific temporary tasks (Faioli, 2017). Anyway, along with the example of policies carried out in Germany, France, Spain, Poland, Italian unions’ position on the best way to protect gig-workers, namely riders, is still building up its own strategy (Don’t GIG Up Report, 2019).

4.2 Collective Agreements and Classification (as to Industry 4.0)

More facts compose the frame concerning the collective bargaining for Industry 4.0 and, specifically, intelligent robotics in the firms’ wall-to-wall perimeters. Such collective bargaining strategy is mainly focused on the in-company mobility and job rotation, job tasks and worker classifications.

Within the Italian system there is a collective pattern which is coupled with a series of legislative developments (first of all Article 8 of Decree Law No. 138/2011, turned into Law No. 148/2011), which have also impacted on the main provision applying to the staff classification system, i.e. Article 2103 of the Italian Civil Code (amended by Decree Law No. 81/2015).

Unlike in the legal systems of other EU countries, in Italy legal regulation has focused on forms of flexibility to be applied when entering (types of contract) or leaving (remedies in case of unfair dismissal) the labour market.

Though new regulations have also dealt with forms of flexibility within the employment contract (i.e. on worker’s in-company mobility), the possibility has strictly given to change (unilaterally or consensually) job tasks in relation to a specific contractual classification system/pay scale framework, even where depending on Industry 4.0.

References

1. Carabelli, U., Spinelli, C.: Riders. Anche il Tribunale di Milano esclude il vincolo di subordinazione. In: *Rivista Giuridica del Lavoro e della Previdenza Sociale*, 1, II, p. 4 ff. (2019)
2. Mengoni, L.: Osservazioni e proposte sulla revisione della legislazione sul rapporto di lavoro, Rapporto approvato dall'Assemblea CNEL, No. 206, 4 June, 1985, available at <https://www.cnel.it/Documenti/Osservazioni-e-Proposte> (1985)
3. Davidov, G.: *A Purposive Approach to Labour Law*, Oxford University Press, Oxford (2016)
4. De Stefano V.: The Rise of the “Just-in-Time Workforce”: On-Demand Work, Crowdwork and Labour Protection in the “Gig-Economy”. In: *Comparative Labour Law & Policy Journal*, 37, p. 471 ff. (2016)
5. Don't GIG Up Report: Directorate-General for Employment, Social Affairs and Inclusion of the European Commission - with Fondazione G. Brodolini (IT), Foundation Institute of Public Affairs-IPA (PL), Istituto delle ricerche economiche e sociali - IRES (FR), UGT Unión General de Trabajadores (ES), UIL (IT), available at http://www.fondazionebrodolini.it/sites/default/files/state_of_the_art_report_0.pdf (2019)
6. Faioli, M.: Jobs App, Gig-Economy e sindacato: In *Rivista Giuridica del Lavoro e della Previdenza Sociale*, 2, I, p. 291 ff. (2017)
7. Faioli, M.: *Mansioni e macchina intelligente*, Giappichelli, Torino (2018)
8. Means, B., Seiner, J.A.: Navigating the Uber Economy. In: *U.C. Davis Law Review*, 49, p. 1511 ff. (2016)
9. Occhino, A.: Per una geografia giuridica del lavoro. In: Occhino, A. (ed.), *Il lavoro e i suoi luoghi*, Vita e Pensiero, Milan (IT), p. 11 ff. (2018)
10. Occhino, A.: Soggettività e nuove rappresentanze nell'economia digitale. In: *Labor*, 1, p. 39 ff. (2019)
11. Prassl, J.: *Humans As Service*, Oxford University Press, Oxford (2017)
12. Rogers, B.: Employment Rights in the Platform Economy: Getting Back to Basics. In: *Harvard Law & Policy Review*, 10, p. 479 ff. (2016)
13. Treu, T., Sacchi, S. (opinions), in *La Repubblica*, 25 June 2017, available at <https://ricerca.repubblica.it/repubblica/archivio/repubblica/2017/06/25/lavoro-trasformare-uber-e-co-in-agenzie-di-ministero18.html> (2017)
14. Stone, K.V.W.: Uber and Arbitration: A Lethal Combination, available at <http://www.epi.org/blog/uber-and-arbitration-a-lethal-combination/> (2016)

Domain knowledge based priors for clustering

Distribuzioni a priori per l'analisi di raggruppamento basate sulla conoscenza di settore

Sally Paganin

Abstract The construction of informative priors based on domain knowledge is a delicate problem, complicated by the fact that the human mind finds it difficult to quantify qualitative knowledge. We focus on the situation in which a prior guess of the data partition is provided, and illustrate how to include such information in a Bayesian mixture model framework. The methodology builds on class of perturbed EPPFs (Exchangeable Partition Probability Function) which centers the prior probability on the most compatible set of partitions, according to the provided guess.

Abstract *La costruzione di distribuzioni a priori informative basate sulla conoscenza di settore è una problematica delicata, complicata dal fatto che la mente umana trova difficoltà nel quantificare le conoscenze di tipo qualitativo. Questo lavoro prende in esame il contesto in cui si ha disposizione una plausibile proposta di partizione dei dati, e illustra come includere tale informazione in modelli di mistura di tipo bayesiano. La metodologia considerata si basa su una classe di EPPF (Exchangeable Partition Probability Function) penalizzate che centrano la distribuzione di probabilità a priori intorno all'insieme di partizioni maggiormente compatibili con la partizione data.*

Key words: Bayesian clustering, centered process, domain knowledge, partition models.

1 Introduction

Mixture models have become increasingly popular tools to model data characterized by the presence of subpopulations, in which each observation belongs to one of a certain number of groups. In particular, observations y_1, \dots, y_N can be divided

Sally Paganin

Department of Statistical Sciences, University of Padova, via Cesare Battisti 241, 35121 Padova,
e-mail: paganin@stat.unipd.it

into $K \leq N$ groups, according to a partition $c = \{B_1, \dots, B_K\}$ with B_k comprising all the indices of data points in cluster k , for $k = 1, \dots, K$. The main underlying assumption of a mixture model is that observations are independent conditional on the partition c and on the vector of unknown parameters $\theta = (\theta_1, \dots, \theta_K)$ indexing the distribution of observations within each cluster. Hence the joint probability density of observations y_1, \dots, y_N can be expressed as

$$p(\mathbf{y}|c, \theta) = \prod_{k=1}^K \prod_{i \in B_k} p(y_i|\theta_k) = \prod_{k=1}^K p(\mathbf{y}_k|\theta_k),$$

with $\mathbf{y}_k = \{y_i\}_{i \in B_k}$ indicating all the observations in cluster k for $k = 1, \dots, K$. In the full Bayesian formulation, a prior distribution is assigned to each possible partition c , leading to a posterior of the form

$$p(c|\mathbf{y}, \theta) \propto p(c) \prod_{k=1}^K p(\mathbf{y}_k|\theta_k).$$

The data partition c is conceived as a random object and elicitation of its prior distribution is a critical issue in Bayesian modeling since the space of all possible partitions grows exponentially fast given its combinatorial nature. Current Bayesian methods often relies on Species Sampling Models (SSM) [7], which avoid dealing with the clustering space directly by inducing a latent partitioning of the data. The induced probability distribution is known in literature as Exchangeable Partition Probability Function (EPPF).

Despite providing tractable tools to deal with mixture models, Bayesian non-parametric priors may be too flexible especially when relevant prior information is available about the clustering, since they lack of a simple way to include this type of information. In particular we focus on the situation in which a base partition c_0 is provided as a prior guess, and we wish to include this information in the prior distribution. To address this problem [6] propose a general strategy to modify a baseline EPPF to shrink the prior probability on partitions towards c_0 . In particular, the prior distribution on all the possible clusterings is defined as proportional to a baseline EPPF multiplied by a penalization term of the type

$$p(c|c_0, \psi) \propto p_0(c) e^{-\psi d(c, c_0)}, \quad (1)$$

with $\psi > 0$ a tuning parameter, $d(c, c_0)$ a suitable distance measuring how far c is from c_0 and $p_0(c)$ indicates a baseline EPPF, that may depend on some parameters. Notice that as $\psi \rightarrow 0$ then $p(c|c_0, \psi)$ corresponds to the baseline EPPF $p_0(c)$, while as $\psi \rightarrow \infty$ then $p(c = c_0) \rightarrow 1$.

The general formulation given in (1) leads to different results on the basis of different choices of EPPF, tuning parameter and distance between partitions. While we refer to [6] for considerations about the choice of EPPFs and tuning parameter, this work focus on characterizing the distance term by providing the definition of a class of a suitable metric between partitions, along with a characterization of

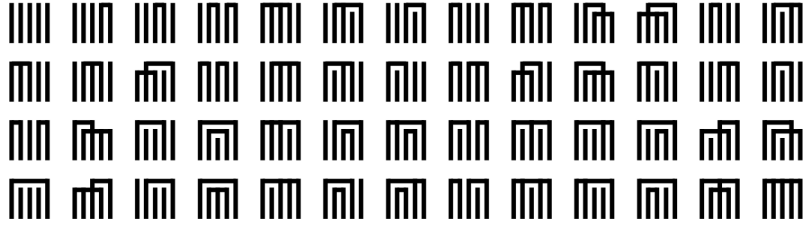


Fig. 1 Genji-mon symbols for all the possible grouping of 5 elements.

neighborhoods induced by the distance. In the following, Section 2 introduces the mathematical definition of set partitions along with concepts derived from lattice theory while Section 3 gives the general definition of distance between partitions. Finally Section 4 provides a characterization of the distance neighborhoods induced by the distance.

2 Set partitions

Let c be a generic clustering of indices $\{1, \dots, N\} = [N]$. It can be either represented as a vector of indices $\{c_1, \dots, c_N\}$ with $c_i \in \{1, \dots, K\}$ for $i = 1, \dots, N$ and $c_i = c_j$ when i and j belong to the same cluster, or as a collection of disjoint subsets (blocks) $\{B_1, B_2, \dots, B_K\}$ where B_k contains all the indices of data points in the k -th cluster and K is the number of clusters in the sample of size N . From a mathematical perspective $c = \{B_1, \dots, B_K\}$ is a combinatorial object known as *set partition* of $[N]$. In denoting a set partition, we either write $\{\{1, 2, 4\}, \{3, 5\}\}$ or $124|35$ using a vertical bar to indicate a break in blocks. By convention, elements are ordered from least to greatest and from left to right within a block; we then order the blocks by their least element from left to right. The collection of all possible set partitions of $[N]$, denoted with Π_N , is known as *partition lattice*. We refer to [8, 1] for an introduction to lattice theory, reporting here some of the base concepts.

According to [4], set partitions seem to have been systematically studied for the first time in Japan (1500 A.D.), due to a parlor game popular in the upper class society known as *genji-ko*; 5 unknown incense were burned and players were asked to identify which of the scents were the same, and which were different. Ceremony masters soon developed symbols to represent all the possible 52 outcomes, so called *genji-mon* represented in Figure 1. Each symbol consists of five vertical bars, with some of them connected by horizontal bars, in correspondence of grouped elements. As an aid to memory, each of the patterns was made after a famous 11th-century novel, *Tales of Genji* by Lady Murasaki, whose original manuscript is now lost, but has made *genji-mon* an integral part of the Japanese culture. In fact, such symbols

continued to be employed as family crests or in Japanese kimono patterns until the early 20th century, and can be found printed in many dresses sold today.

First results in combinatorics focused on enumerating the elements of the space, making their appearance during the 17th century, still in Japan. For example, the number of ways to assign N elements to a fixed number of K groups is described by the *Stirling number of the second kind*

$$\mathcal{S}_{N,K} = \frac{1}{K!} \sum_{j=0}^K (-1)^j \binom{K}{j} (K-j)^N,$$

while the *Bell number* $\mathcal{B}_N = \sum_{K=1}^N \mathcal{S}_{N,K}$ describes the number of all possible set partitions of N elements. Refer to [4] for more information on history and algorithms related to set partitions and other combinatorial objects.

2.1 Poset representation of partition lattice

The interest progressively shift from counting elements of the space to characterizing the structure of space partitions using the notion of partial order. Consider Π_N endowed with the set containment relation \leq , meaning that for $c = \{B_1, \dots, B_K\}, c' = \{B'_1, \dots, B'_{K'}\}$ belonging to Π_N , $c \leq c'$ if for all $i = 1, \dots, K, B_i \subseteq B'_j$ for some $j \in \{1, \dots, K'\}$. Then the space (Π_N, \leq) is a *partially ordered set* (poset), which satisfies the following properties:

1. Reflexivity: for every $c \in \Pi_N$, $c \leq c$,
2. Antisymmetry: if $c \leq c'$ and $c' \leq c$, then $c = c'$,
3. Transitivity: if $c \leq c'$ and $c' \leq c''$, then $c \leq c''$.

Let $<$ be the relation on Π_N such that $c < c'$ if and only if $c \leq c'$ and $c \neq c'$. For any $c, c' \in \Pi_N$, it is said that c is *covered* (or *refined*) by c' if $c \leq c'$ and there is no c'' such that $c < c'' < c'$ and indicate with $c \prec c'$ such relation. This covering relation allows one to represent the space of partitions by means of the *Hasse diagram*, in which the elements of Π_N correspond to nodes in a graph and a line is drawn from c to c' when $c \prec c'$; in other words, there is a connection from a partition c to another one when the second can be obtained from the first by splitting or merging one of the blocks in c . See Figure 2 for an example of Hasse diagram of Π_4 . If two elements are not connected, as for example partitions $\{1, 2\}\{3, 4\}$ and $\{1, 3\}\{2, 4\}$, they are said to be *incomparable*. Conventionally the partition with just one cluster is represented at the top of the diagram and denoted as 1, while the partition having every observation in its own cluster at the bottom and indicated with 0.

The space Π_N is also a *lattice*, for the fact that every pair of elements has a *greatest lower bound* (g.l.b.) and a *least upper bound* (l.u.b.) indicated with the “meet” \wedge and the “join” \vee operators, i.e. $c \wedge c' = \text{g.l.b.}(c, c')$ and $c \vee c' = \text{l.u.b.}(c, c')$ and equality holds under a permutation of the cluster labels. An element $c \in \Pi_N$ is an upper bound for a subset $S \subseteq \Pi_N$ if $s \leq c$ for all $s \in S$, and it is the least

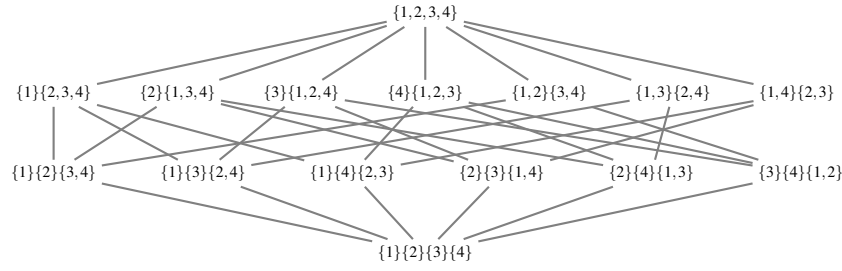


Fig. 2 Hasse diagram for the lattice of set partitions of 4 elements. A line is drawn when two partitions have a covering relation. For example $\{1\}\{2,3,4\}$ is connected with 3 partitions obtained by splitting the block $\{2,3,4\}$ in every possible way, and partition **1** obtained by merging the two clusters.

upper bound for a subset $S \subseteq \Pi_N$ if c is an upper bound for S and $c \leq c'$ for all upper bounds c' of S . The lower bound and the greatest lower bound are defined similarly, and the definition applies also to the elements of the space I_N . Consider as an example $c = \{1\}\{2,3,4\}, c' = \{3\}\{1,2,4\}$; their greatest lower bound (g.l.b.) is $c \wedge c' = \{1\}\{3\}\{2,4\}$ while the least upper bound (l.u.b.) is $c \vee c' = \{1,2,3,4\}$. Looking at the Hasse diagram in Fig 2 the g.l.b. and l.u.b. are in general the two partitions which reach both c and c' through the shortest path, respectively from below and from above.

3 Distances on the partition lattice

The representation of the space of set partitions Π_N from lattice theory, provides a useful framework to define metrics between partitions. In fact, the distance between any two partitions can be defined by means of the Hasse diagram as the length of any shortest path between them, which necessarily passes through the meet or join of two partitions.

More general distances arise when the graph is weighted, meaning that every edge is associated with a strictly positive weight; then the distance between any two elements is the weight of the lightest path between them, where the weight of a path is the sum over its edges of their weight. Weights over the edges of the Hasse diagram are usually defined starting from a function v on the lattice Π_N having the following properties.

Definition 1. A lattice function $v : \Pi_N \rightarrow \mathbb{R}^+$, is said to be

- *strictly order-preserving* if $v(c) > v(c')$, for $c, c' \in \Pi_N$ such that $c > c'$.
- *strictly order-reversing* if $v(c) > v(c')$, for $c, c' \in \Pi_N$ such that $c < c'$.
- *supermodular* if $v(c \vee c') + v(c \wedge c') - v(c) - v(c') \geq 0$, for any $c, c' \in \Pi_N$.
- *submodular* if $v(c \vee c') + v(c \wedge c') - v(c) - v(c') \leq 0$, for any $c, c' \in \Pi_N$.

We report here a useful result from lattice theory referring to [3] and [2]. Given a lattice function v weights w_v on edges between $\{c, c'\}$ are defined as

$$w_v(\{c, c'\}) = |v(c) - v(c')|,$$

with distance between two partitions being the minimum- v -weighted path. Properties outlined in Definition 1 guarantee that such path visits either the meet or the join of any two incomparable partitions; which one of the two depends on whether the function v is supermodular or submodular.

Proposition 1. *For any strictly order-preserving (order-reversing) function v , if v is supermodular, the minimum- v -weight partition distance is*

$$d_v(c, c') = v(c) + v(c') - 2v(c \wedge c') \quad (d_v(c, c') = v(c) + v(c') - 2v(c \vee c')),$$

while if v is submodular

$$d_v(c, c') = 2v(c \vee c') - v(c) - v(c') \quad (d_v(c, c') = 2v(c \wedge c') - v(c) - v(c')).$$

4 Distance neighborhoods on the partition lattice

Due to the discrete nature of the space of partition, the distance $d_v(c, c_0)$ takes a finite number of discrete values $\Delta = \{\delta_0, \dots, \delta_L\}$, with L depending on c_0 and on the distance $d(\cdot, \cdot)$. We can define distance neighborhoods as

$$s_l(c_0) = \{c \in \Pi_N : d_v(c, c_0) = \delta_l\}, \quad l = 0, 1, \dots, L, \quad (2)$$

hence sets of partitions having the same fixed distance from c_0 . For $\delta_0 = 0$, $s_0(c_0)$ denotes the set of partitions equal to the base one, meaning that they differ from c_0 only by a permutation of the cluster labels. Then $s_1(c_0)$ denotes the set of partitions with minimum distance δ_1 from c_0 , $s_2(c_0)$ the set of partitions with the second minimum distance δ_2 from c_0 and so on. Hence the exponential term in (1) penalizes equally partitions in the same set $s_l(c_0)$ for a given δ_l .

A trivial example can be obtained by considering the rank function, i.e. $r(\cdot) : \Pi_N \rightarrow \mathbb{Z}^+$ such that $r(c) = N - |c|$, which is a strictly order-preserving lattice function. For example, considering partitions in the Hasse diagram in Figure 2, the rank of the bottom partition 0 is equal to 0 and increases by 1 for each level of the graph up to 3 for top partition 1. Then the minimum-rank-weighted distance can be computed as $d_r(c, c') = 2r(c \vee c') - r(c) - r(c')$, since the function is also submodular. Notice that the rank assigns to every edge between partitions a unit weight, and then d_r is indeed the shortest path distance.

Figure 3 provides a representation of the distance neighborhoods as defined in 2 induced by the rank function when the base partition corresponds to $c_0 =$

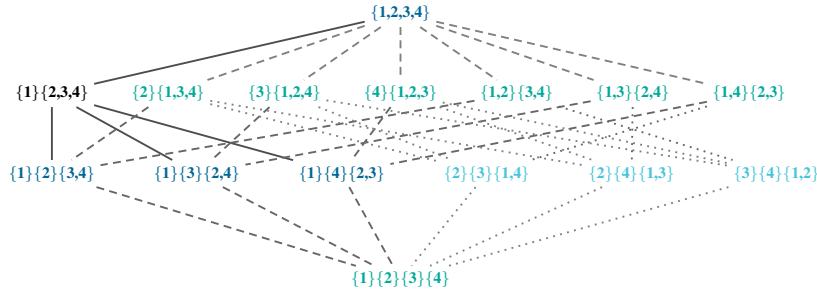


Fig. 3 Representation of distance neighborhoods on the poset lattice Π_4 for $c_0 = \{1\}\{2,3,4\}$ when the chosen distance is induced from the rank function. Partitions are colored from the closest to the most distant according to dark-light gradient.

$\{1\}\{2,3,4\}$. It can be notice that the closest partitions, besides c_0 itself, in $s_1(c_0)$ are the ones obtained with a single operation of split or merge on c_0 , while the second closest ones by applying another operation of split or merge on the partitions in $s_1(c_0)$, and so on. This kind of behavior can be observed for all functions which induce unit weight on the edges of the partition lattice.

Another important measure of distance between two partitions, is the Variation of Information (VI), introduced axiomatically in information theory [5], which also belongs to the class of distances derived from a lattice function. In particular, consider the Shannon entropy $H(\cdot) : \Pi_N \rightarrow \mathbb{R}^+$ defined as $H(c) = -\sum_{i=1}^K |B_i|/N \log_2(|B_i|/N)$ which is a submodular and strictly order-reversing function, hence inducing distance

$$d_H(c, c') = VI(c, c') = 2H(c \wedge c') - H(c) - H(c'). \quad (3)$$

The VI ranges on a finite subset in $[0, \log_2 N]$, and in this case the weights assigned to the edges differs from the unit weight, leading to finer characterization of the distance neighborhoods as it can be seen from Figure 4. In general the closest partitions are the ones which differs from c_0 by merging two singleton clusters or splitting a cluster of size two into singletons. If neither is possible, the closest partitions differs from c_0 by a split operation on the smallest cluster of size k into a singleton and a cluster of size $k - 1$ or, as in the example, by a merge operations on these last two clusters.

References

1. Davey, B. A. and Priestley, H. A.: Introduction to Lattices and Order. Cambridge university press. (2002)
2. Deza, M. M. and Deza, E.: Encyclopedia of Distances. Springer Berlin Heidelberg. (2009)
3. Grätzer, G.: General Lattice Theory. Springer Science & Business Media. (2002)
4. Knuth, D.: The Art of Computer Programming: Generating All Trees. History of Combinatorial Generation. Addison-Wesley. (2006)

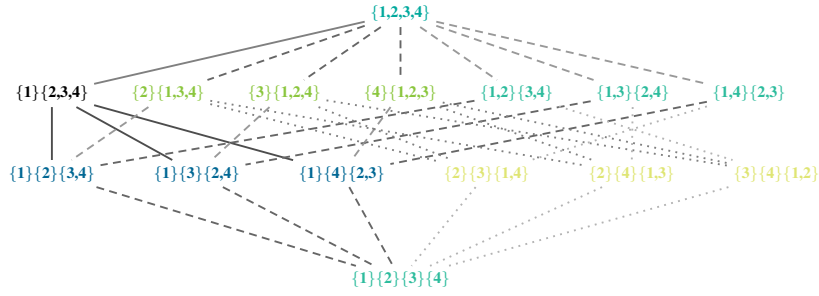


Fig. 4 Representation of distance neighborhoods on the poset lattice Π_4 for $c_0 = \{1\}\{2,3,4\}$ when the chosen distance is induced from the Shannon entropy function. Partitions are colored from the closest to the most distant according to dark-light gradient.

5. Meilä, M.: Comparing clusterings - an information based distance. *Journal of Multivariate Analysis* **98**(5), 873 – 895. (2007)
6. Paganin, S., Herring, A. H, Olshan, A. F., Dunson, D. B. and The National Birth Defect Study: Centered Partition Process: Informative Priors for Clustering. *arXiv preprint* arXiv:1901.10225. (2018)
7. Pitman J.: Exchangeable and partially exchangeable random partitions. *Probability theory and related fields* **102**, 145–158. (1995)
8. Stanley, R. P.: Enumerative Combinatorics. Vol. 1. Cambridge University Press. (1997)

Clustering of Behavioral Spatial Trajectories in Neuropsychological Assessment

Analisi dei gruppi di traiettorie spaziali nella valutazione neuropsicologica

Francesco Palumbo, Antonio Cerrato, Michela Ponticorvo, Onofrio Gigliotta, Paolo Bartolomeo, Orazio Miglino

Abstract Neuropsychological assessment consists in the evaluation of cognitive skills based on subjects' performance. Here we present an integrated procedure based on an enhanced version of a spatial cognition task, the Baking Tray Task (BTT). In this version, data are automatically recorded and immediately available for the analysis. We then develop a two-step statistical analysis to identify groups of homogeneous patterns of performance: (1) a functional principal components analysis represents the data as points on a subspace; (2) a finite mixture model-based clustering algorithm detects groups having similar patterns of performance. We present the results from a random sample of 114 healthy subjects. Future research will aim to offer diagnostic procedures for spatial deficits, based on these methods.

Abstract *La valutazione neuropsicologica è basata sulle valutazioni di abilità cognitive, come le abilità spaziali. Questo studio presenta una procedura integrata che si basa su una versione evoluta del Baking Tray Task (BTT), in cui i dati vengono acquisiti automaticamente e resi subito disponibili per l'analisi. L'analisi statistica mira ad individuare gruppi di comportamenti spaziali omogenei e si articola su due passi: (1) l'analisi delle componenti principali funzionali, che consente di rappresentare i dati come punti in un sottospazio; (2) un algoritmo di clustering basato su un modello di mistura per trovare gruppi di comportamenti simili. Presentiamo i risultati ottenuti su un campione casuale di 114 soggetti sani. In futuro, ci proponiamo di offrire procedure di diagnosi di deficit spaziali basate su questi metodi.*

Key words: Spatial Behavior; Clustering; Functional Data Analysis; Neuropsychology; Assessment

F. Palumbo, A. Cerrato, M. Ponticorvo, O. Gigliotta, O. Miglino
University of Naples Federico II Italy
e-mail: {fpalumbo, antonio.cerrato, michela.ponticorvo, onofrio.gigliotta, miglino}@unina.it

P. Bartolomeo
Inserm U 1127, CNRS UMR 7225, Sorbonne Université, Institut du Cerveau et de la Moelle épinière, ICM, Hôpital de la Pitié-Salpêtrière, Paris, France
e-mail: paolo.bartolomeo@icm-institute.org

1 Introduction

Neuropsychological assessment consists in the evaluation of cognitive skills based on subjects' performance. A class of cognitive skills is represented by the spatial abilities, which subtend spatial cognition and interaction with external stimuli. Spatial tests are important to diagnose and manage deficits of spatial abilities, which are a substantial source of handicap in brain-damaged patients [3]. Most spatial tests are based on traditional "paper-and-pencil" methods. Paper-and-pencil testing is usually easy to administer, but is also time- and labor-consuming for examiners and participants [1]. Furthermore, the test results are not usually immediately available in numerical form for diagnostic purposes.

Typical paper-and-pencil tests for spatial abilities are target cancellation tasks, [3], in which patients are required to find and cancel target stimuli presented on a sheet of paper; this class of tests evaluates visual scanning and search abilities. It has also been administered to artificial agents [17, 11, 12].

More recently, efforts are made to develop more ecological forms of assessment, simulating real-life scenarios [9].

One such spatial test is the Baking Tray Task (BTT) test [19], whereby participants are requested to dispose as evenly as possible 16 cubes on a board.

Here we propose a technology-enhanced version of the BTT. In the present version, the spatial coordinates of the disposed buns are automatically acquired, and made immediately available for statistical analysis.

2 Methods

The BTT was originally devised by Tham e Tegner [19] as a test of visual neglect. Neglect is a disabling condition mainly occurring in patients with strokes in the right hemisphere. Neglect patients ignore left-sided objects, thus living in a virtually halved world, and may even neglect their own left limbs [3].

Participants to the BTT are requested to dispose 16 cubes on a board, as evenly as possible as if they were buns to be put in the oven. It is usually scored by calculating the difference between the number of cubes disposed in the left and right halves of the boards. A difference greater than two is considered diagnostic of visual neglect. The BTT is considered to be a sensitive test for screening purposes and longitudinal studies. It seems able to detect all cases of at least moderately severe visual neglect, whereas standard neglect tests can miss some of the cases [13]. Another advantage of the BTT is that it seems to be relatively free of practice effects, because of its low requirements on attention and memory.

We developed a new version of the BTT, by exploiting the augmented reality and artificial vision technologies [10, 6], based on tangible interfaces [16, 7]. The participants' task is strictly the same, but for each subject, the enhanced version continuously registers the position of the disposed objects, and generates an ordered sequence of 16 pairs of coordinates in the X, Y Cartesian space. The new tool fea-

tures an easy, fast and inexpensive way to collect and storing the data. It represents a further development of a previous prototype [4, 5].

The test was administrated to 114 healthy participants (61 female, 26.8 mean age, 7.3 sd) in strict accordance with the Tham and Tegner's guidelines. The only differences with the original version are that the current version uses tagged discs rather than cubes, and a surface of smaller dimensions. These adjustments were already proposed in previous works [2, 8]. The recorded data can easily be exported for analysis.

3 Results and Discussion

Each record consists of sixteen pairs of Cartesian coordinates that correspond to an equal number of points on the board. Taking into account the disposing order, each participant's performance describes a trajectory. Figure 1 shows an example of a experimental session. Figure 1(a) represents the registered positions of the sixteen discs placed on the board that are numbered according to the placing order; Figure 1(b) represents the X and the Y coordinates, independently.

Given the importance of left-right coordinates for neglect, the X coordinates are the most relevant for our scope; however, Y coordinates can reveal additional information about the participants' performance (e.g., altitudinal, or vertical neglect).

The sixteen X coordinates are assumed as points belonging to an unknown function defined on the time domain T with $t = 0, \dots, 15$. Each function is approximated through eight basis splines of order 2. The natural choice of the number of bases would be 16, yet empirical evidence proved that the eight basis approximation ensures good approximation and a reduced number of parameters [14, Chap. 5]. The same pre-treatment was applied to the Y coordinates too. However, results are not illustrated in the present article. The statistical analysis on the X coordinates consists of two steps: (i) data reduction through the principal component analysis (PCA) on

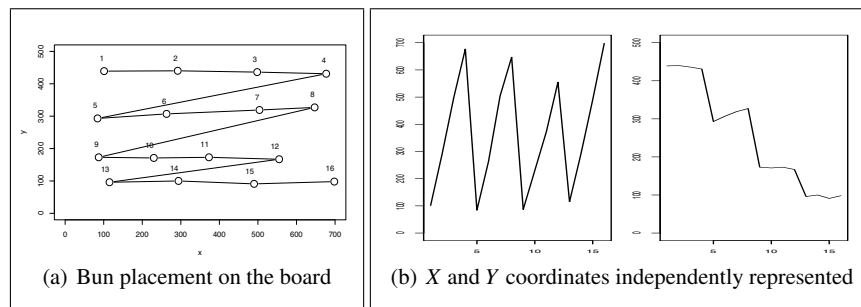


Fig. 1: Example of the BTT recording

the functional data [18, Chap. 18]; (ii) mixture model data clustering on the reduced subspace to identify groups of similar behaviors [15].

Let p variables have variance-covariance matrix Σ , the aim of principal component analysis (PCA) is to find new variables that are linear combinations of the p observed variables, so that they have maximum variability and are orthogonal. A vector \mathbf{a}_1 (with $\mathbf{a}_1^\top \mathbf{a}_1 = 1$) of length p is found to maximize the variance $\mathbf{a}_1^\top \Sigma \mathbf{a}_1$. So that, the first principal component $\mathbf{z}_1 = \mathbf{X} \mathbf{a}_1$ is a linear combination of the variables that has the maximum variance. Further principal components are derived such that each has maximum variance given that it is orthogonal to the previous ones.

The main goal of PCA for functional data is essentially the same, yet the different nature of the data implies some changes. Let \mathbf{Y} be the generic $n \times T$ functional data matrix and $y_i = [y_i(0), y_i(1), \dots, y_i(t), \dots, y_i(T)]$ a generic row vector in \mathbb{R}^T where $y_i(t) = f(t)$ indicates the value of the unknown function $f(\cdot)$ in t . Without loss of generality we can assume that $y(t)$ has the same origin and scale for any $t \in [0, T]$, then the cross-product function

$$c(t, t') = \frac{1}{n} \sum_i y_i(t) y_i(t') \quad (1)$$

has equivalent informative content of the Pearson's correlation index in the functional data analysis (FDA) framework. Under these premises, it is intuitive that in functional PCA to calculate the first *principal component scores* f_1 consists in finding the vector ξ_1 that maximizes the quantity $\sum_{i1} f_{i1}^2$, where $f_{i1} = \int \xi_1(t) y_i(t) dt$, under the usual constraint $\|\xi_1\|^2 = 1$. Then, the second maximum variability direction consists in computing the vector ξ_2 that maximizes $\sum_i f_{i2}^2(t)$, under the constraints: $\|\xi_2\|^2 = 1$ and $\xi_1^\top \xi_2 = 0$ (equivalent to the orthonormality constraint in the PCA).

However, to rewrite the above maximization problem in term of eigenanalysis to compute the *principal component scores*, it is necessary to replace the vector $y_i(t)$ with additional variables, which are transformation of $y_i(t)$, to use the PCA. A more flexible solution that ensures better approximation is represented by the b-spline functions. Given a set of M points called *knots*, with $m = 1, \dots, M$, a b-splines is piecewise polynomial continuous at each knot. It is worth to remark that the set of linear splines with fixed knots is still a vector space or a basis in T . The b-splines allows for efficient computations even when the number of knots M is large. The observed function $y(t)$ can be expanded in terms of a vector $\phi(t)$ of M basis functions

$$y(t) = \mathbf{C} \phi(t) \quad (2)$$

analogously the generic eigenfunction becomes $\xi_j(t') = \mathbf{b}^\top \phi(t')$. Replacing the 2 in 1 the crossproduct becomes

$$c(t, t') = \frac{1}{n} \phi^\top(t) \mathbf{C}^\top \mathbf{C} \phi(t'). \quad (3)$$

The FPCA then becomes

$$\frac{1}{n}\phi^\top(t)\mathbf{C}^\top\mathbf{C}\int\phi(t')\phi^\top(t')dt'\mathbf{b}_j=\lambda_j\phi^\top(t')\mathbf{b}_j, \quad (4)$$

replacing $\mathbf{J} = \int\phi(t')\phi^\top(t')dt'$ the 4 becomes

$$\frac{1}{n}\phi^\top(t)\mathbf{C}^\top\mathbf{C}\mathbf{J}\mathbf{b}_j=\lambda_j\phi^\top(t)\mathbf{b}_j, \quad (5)$$

under the constraint $\mathbf{b}_j^\top\mathbf{J}\mathbf{b}_j = 1$.

Technically speaking the quantity $\phi^\top(t)\mathbf{b}_j$ is called *harmonic* and is an averaged approximation of the observed functions $y_i(t)$. Exploiting the usual properties of the PCA, each function $y_i(t)$ can be represented by a point into the orthogonal subspace spanned by the vectors \mathbf{b}_j , where $j = 1, \dots, J$. The choice of J follows the usual rules of the PCA, in this proposal it has been set $J = 2$.

The orthogonal subspace spanned by the vectors \mathbf{b}_j offers many opportunities for further analyses.

In the present study, we used a very flexible model based clustering approach that is based on the finite Mixture of Generalized Hyperbolic Distributions (MGHD). In short, a univariate or multivariate finite mixture distribution is a (probability) density function defined as the weighted sum of K (with $K \geq 2$) density functions of the same family that differ in at least one parameter. Each density function of the mixture is called a component. The best known and probably the most widely used is the finite mixture of Gaussian distributions, albeit mixtures can be defined for any distribution family, at least from a theoretical point of view. Model-based clustering assumes that data were generated from a mixture of K components, where K is given, and that each group refers to one component of the mixture. Model-based clustering algorithms estimate the parameters (or the parameter vectors) of each component and a set of weights $\pi_1, \pi_2, \dots, \pi_K$, such that $\pi_k > 0$ for any $k \in 1, \dots, K$ and $\sum_k \pi_k = 1$. One important property of the generalized hyperbolic random variable is that it can be generated combining a generalized inverse Gaussian and a multivariate Gaussian making feasible the parameter estimation of the corresponding density function. The MGHD is not the only possible solution to find the groups in the data, of course. Other approaches can lead to similar results. However, we believe that the flexibility of the MGHD helps to obtain easier and more clear interpretations of the data.

The following results are based on the information provided by coordinates with respect to the horizontal axis; upward or downward shifts are not taken into account. Such extra information (which is not considered in the original BTT, but the present, enhanced version can provide for future analyses) might better define the clinical profile.

The plot in figure 2 represents the b-splines approximated x functions with respect to the first two harmonics obtained by the functional PCA. It also shows the clustering results representing the three groups and the corresponding bivariate density mixture. Two homogeneous groups emerged (circles and plus symbols, on the left and the right of the plot). A third, more heterogeneous group was also present

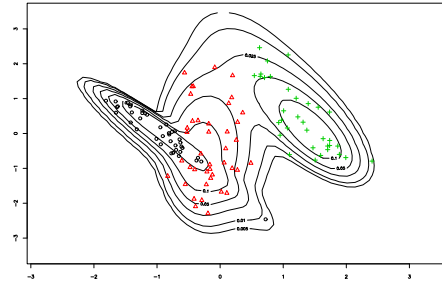


Fig. 2: Functional PCA (h1 and h2) with the finite mixture density of generalized hyperbolic distributions that identifies the three groups.

(triangles at the center of the plot). The interpretation is similar to the classical PCA, with the position in correspondence of the axes origin indicates average profiles.

We then plotted the three groups as a function of participants' gender (Figure 3). Each pair of plots refers to each of the three groups, and represents the x functions and the corresponding projections on the principal subspace spanned by the first

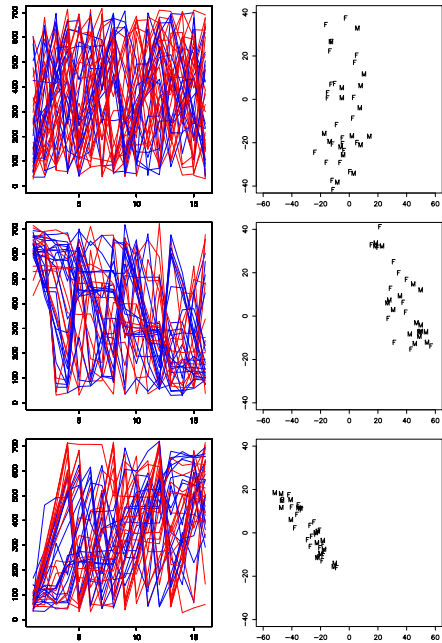


Fig. 3: Functional PCA (h1 and h2) with the finite mixture density of generalized hyperbolic distributions that identifies the three groups. Red, female participants (F); blue, male participants (M).

two harmonics. Statistical units are represented by the symbols "F" (for female) and "M" (for male).

The analysis of the frequencies distribution of male and female within groups reveals a weak and not statistically significant association, as it is confirmed by the chi-square test for the independence ($\chi^2 = 3.13$, $p\text{-value} = .21$, $df = 2$). The first pair of plots from top in figure 3 refers to the most heterogeneous group ($n = 41$, $F=61\%$) that is located in correspondence of the origin of the axes. These subjects started to place the buns from several different positions, following no clear plan. The second pair corresponds to those subjects ($n = 34$, $F=41\%$) that mainly started from the right side and then went to the left according to a periodic movement (they came back to the right side after four placed buns on average). The left-sided plot shows the moderate masculine predominance (blue lines) in this group. The third plots represent participants who tended to start from the left side ($n = 39$, $F=56\%$), and showed a feminine predominance (red lines in the left-sided plot).

4 Conclusion

We presented a technology-enhanced version of the BTT, a sensitive test of spatial cognition. Advanced analysis of the results obtained in a neurotypical population showed evidence of three main patterns of performance, with possible gender differences. The establishment of typical performance patterns on the BTT will hopefully lead to the development of precise diagnostic procedures for spatial deficits in brain-damaged patients.

References

1. Appellos, P., Karlsson, G., Thorwalls, A., Tham, K., Nydevik, I.: Unilateral neglect: further validation of the baking tray task. *Journal of Rehabilitation Medicine* **36**(6), 258–261 (2004)
2. Bailey, M.J., Riddoch, M.J., Crome, P.: Test–retest stability of three tests for unilateral visual neglect in patients with stroke: Star cancellation, line bisection, and the baking tray task. *Neuropsychological Rehabilitation* **14**(4), 403–419 (2004)
3. Bartolomeo, P.: Attention disorders after right brain damage: Living in halved worlds. Springer, (2013)
4. Cerrato, A., Ponticorvo, M.: Enhancing neuropsychological testing with gamification and tangible interfaces: The baking tray task. In: *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pp. 147–156. Springer (2017)
5. Cerrato, A., Ponticorvo, M., Bartolomeo, P., Miglino, O.: Btt-scan: An ecological and technology enhanced tool to assess visual neglect. In: *Cognitive Processing*, vol. 19, pp. S36–S36. Springer, Heidelberg (2018)
6. Cerrato, A., Siano, G., De Marco, A.: Augmented reality: from education and training applications to assessment procedures. *Qwerty-Open and Interdisciplinary Journal of Technology, Culture and Education* **13**(1) (2018)
7. Di Fuccio, R., Ponticorvo, M., Ferrara, F., Miglino, O.: Digital and multisensory storytelling: narration with smell, taste and touch. In: *Smart Education and e-Learning*, pp. 329–338. Springer (2016)

8. Facchin, A., Beschin, N., Pisano, A., Reverberi, C.: Normative data for distal line bisection and baking tray task. *Neurological Sciences* **37**(9), 1531–1536 (2016)
9. Franzen, M.D., Wilhelm, K.L.: Conceptual foundations of ecological validity in neuropsychological assessment. (1996)
10. Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F.J., Marín-Jiménez, M.J.: Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition* **47**(6), 2280–2292 (2014)
11. Gigliotta, O., Bartolomeo, P., Miglino, O.: Approaching neuropsychological tasks through adaptive neurorobots. *Connection Science* **27**(2), 153–163 (2015).
12. Gigliotta, O., Seidel Malkinson, T., Miglino, O., Bartolomeo, P.: Pseudoneglect in visual search: Behavioral evidence and connectional constraints in simulated neural circuitry. *eNeuro* **4**(6) (2017).
13. Halligan, P.W., Marshall, J.C.: Left visuo-spatial neglect: A meaningless entity? *Cortex* **28**(4), 525–535 (1992)
14. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer (2011)
15. McNicholas, P.D.: Mixture model-based classification. Chapman and Hall/CRC (2016)
16. Miglino, O., Ponticorvo, M.: Enhancing multi-sensory and handling-based psychopedagogical approaches through new technologies. *PROGEDIT*, Bari, Italy (2018)
17. Pacella, D., Ponticorvo, M., Gigliotta, O., Miglino, O.: Basic emotions and adaptation. a computational and evolutionary model. *PLOS ONE* **12**, 1–20 (2017).
18. Ramsay, J.O., Silverman, B.W.: Principal components analysis for functional data, pp. 147–172. Springer New York, New York, NY (2005).
19. Tham, K., Tegner, R.: The baking tray task: a test of spatial neglect. *Neuropsychological Rehabilitation* **6**(1), 19–26 (1996)

What is wrong in the debate about smart contracts

Smart contract e diritto: riflessioni critiche su un dualismo fuorviante

Roberto Pardolesi and Antonio Davola

Abstract In recent years, the debate about smart contracts, the blockchain, and their interaction with the law has been constantly intensifying, resulting in a vast multiplicity of contributions that try to deal with these new technologies, the role they are likely to assume in the society, and their impact on the legal framework. In our investigation, we first address the main shortcomings, and conceptual errors in the traditional approach to the legal analysis of smart contracts; then, we ultimately offer a critical perspective of how the debate should develop in the future to properly address these technologic challenges.

Abstract *Il dibattito in materia di smart contract, blockchain, e sulla relazione che queste tecnologie intrattengono con le categorie giuridiche tradizionali ha vissuto – in anni recenti – una crescente attenzione: un numero sempre maggiore di contributi ha cercato di delineare il ruolo che queste innovazioni si prestano a giocare nella società contemporanea, nonché il loro impatto sul quadro normativo esistente. Il contributo si propone, innanzitutto, di evidenziare le principali problematicità e gli errori concettuali che caratterizzano l'approccio tradizionale all'analisi giuridica degli smart contract; sulla base di tali riflessioni, successivamente, il tema viene sviluppato in prospettiva critica al fine di fornire alcune indicazioni in merito a come, auspicabilmente, l'analisi del tema dovrebbe svilupparsi in futuro al fine di confrontarsi in maniera adeguata con le sfide che queste tecnologie emergenti pongono.*

Key words: Smart Contract; Blockchain; Law&Technology; Regulation; Contract Law; Private Law.

¹

Roberto Pardolesi, Università LUISS G. Carli, Roma; email: rpardole@luiss.it

Antonio Davola, Istituto S. Anna, Pisa; email: a.davola.sssuo@gmail.com

1 The future is upon us...isn't it?

Around two dozen years ago a computer scientist named Nicholas J. Szabo, in an attempt to canvass the legal implications of his techno-approach, coined the – unexpectedly successful – label “Smart Contracts”, which is nowadays adopted to address a vast range of ‘supercontracts’ that, like mighty superheroes, are supposed to handle any task by themselves.

Truth to be told, Szabo’s proposal was more discrete than what the subsequent debate would suggest: he defined smart contracts as “a set of promises, specified in digital form, including protocols within which the parties perform on these promises”,¹ offering a preliminary overview of their potential applications. Consistently with this view, his initial overview assumed as paradigmatic reference the typical “vending machine” – automatically providing goods and services after receiving an adequate payment. Despite this somewhat humble incipit, the debate over smart contracts rapidly moved towards more ambitious frontiers: from being mere dispensers of services, smart contracts evolved firstly into isomorphic security protocols, designed to ensure compliance with the terms of an agreement, and then even further into multilateral, “fraud-proof”, protocols.² The ultimate step of this Darwinian-like path is the – widely diffused – current belief that smart contracts are tools capable to subsume any aspect of an agreement within the technological realm, consistently with (a misguided reading of) Lessig’s lesson stating that, nowadays, “code is law”. This perspective is the result of a mixture of (small) concrete developments and (several) utopistic hypotheses; yet it led legal scholars to evaluate the possibility to create “contractware”, feeding informatic devices with the terms and clauses of the agreement and ultimately allowing a code to operate as a gatekeeper for the contract, being able to guarantee *ex ante* the fulfilment of the obligations necessary for its diligent execution.

The theoretical framework has been further alimented by the subsequent technological advancements and, in particular, by the entrenched belief that artificial intelligence will, sooner or later, replace any human activity. A.I. is nowadays seen as fit to operate also in those fields where interactions are not predetermined, encompassing systemic uncertainties that can be solved only through decisional synthesis. The debate over silicon-based contracts is further stimulated by the developments of cryptographic technologies: the diffusion of the (much fabled) blockchain,³ operating as a perfect register through its decentralized ledgers, and the crescent attention towards its application as a platform for the circulation of information and goods (such as bitcoins), attracted scholars’ growing interest, intellectual efforts and monetary speculations. Blockchain intertwined with the very own definition of smart contracts, and currently –as will be seen in a

¹ Smart Contracts: Building Blocks for Digital Markets, <https://perma.cc/YC35-2MXQ>, 1996.

² Even if, currently, violations of electronic agreements – especially the ones involving cryptocurrencies – are still present and frequent. See e.g. U.S. RODRIGUES, Law and the Blockchain, 104 Iowa L. Rev. 679, 695 ff. (2018).

³ For a comprehensive overview of the topic, and its legal applications, see (forthcoming) P. HACKER, I. LIANOS, G. DIMITROPOULOS and E EICH, Regulating Blockchain: Techno-Social and Legal Challenges, Oxford University Press, www.ssrn.com, November 2018.

What is wrong in the debate about smart contracts

while- the vast majority of legal contributions in the field cannot discuss one of the two notions, without addressing the other one as well. Whether this tendency is correct or not is an aspect that we will address in the course of this work. As of now, another aspect must be added to the scene: legal scholars – acting as sort of newly-formed engineers – put another layer of complexity on the field, and questioned the very characteristics and meaning of the notion of contract, its underlying principles and legal regime. Enthusiasts of technologies began to wonder whether the rise of smart contracts was an undisputable symptom of the death of the traditional one, arguing in favor of a complete and needed revolution of the role and functions of the legal professions.

Before going more in depth of these considerations, we want to reassure our readers: in this work, it is not our intention to indulge in futuristic speculations; furthermore, you will not find in this article any umpteenth explanation of how the blockchain– and its block system – works, can foster reliability, transparency, stability, self-enforcement etc., both within and outside the legal realm. Our (modest) goal is to highlight the significant amount of contradictions characterizing the current debate over smart contracts (probably not so smart and not even able to express a contract identity). Then, some critical remarks on their aptitude to revolutionize traditional contract law will follow, in order to evaluate whether the legal implications of these changes are (at least) as relevant as the technological ones.

2 The everlasting pursuit of contract 2.0

Before we get to illustrate the reasons behind our skepticism toward the (alleged) disruptive effects of smart contracts for the legal framework, it is opportune to offer a (hopefully concise) overview of the relevant literature. This is necessary, as a matter of fact, to provide a solid basis for our considerations, and to contextualize our reflections within the current state of the art of the legal debate.

In conducting this activity, some preliminary caveat are to be taken into account: the amount of contribution addressing the topic – especially when interdisciplinary works are considered – is vast and heterogeneous. As a consequence, we restricted the scope of our analysis to contributions targeting the specific problem of smart contracts from the perspective of law and regulation. And yet, even after engaging into a thoughtfully selective process, excluding (on one hand) works by technicians generically dealing with regulatory topics and (on the other one) contributions by legal scholars misinterpreting the inner characteristics of the technology, the task proved itself arduous. Note that this is another significant proof of how difficult it is to develop a true interdisciplinary approach to the topic, suggesting – once again – caution for those who want to engage in the area of Law&Technology.

The manifold framework of contributions in the field unveils the presence of purely theoretical speculations, next to works actually engaging with (when not

approximately evoking) the legal practice. A common feature that many of these researches share is the goal of providing a definition of smart contracts. The goal is often met through extensive – when not exhausting – overviews of all the technicalities surrounding the blockchain and its derivative structures (in primis the cryptocurrencies and the Ethereum platform). Then, authors usually provide considerations on the (casual or structural) relation between the abovementioned elements and smart contracts, intended as allegedly legal phenomena.¹

Regarding this last aspect, part of the literature generally qualifies smart contracts as automatized systems (bot contracts?) for the execution of contractual agreements: such a definition is sufficiently vague to encompass both high-frequency financial transactions or (potentially collusive?) interactions amongst algorithms implemented in the setting of prices to reach profit maximization. In any case, no human interaction is ever present, and the whole activity is conducted by the smart code on its own. According to this wide notion, segmented fragments of an agreement can be considered smart contracts as well: even when no contractual intention is present, they are nevertheless instrumental to the realization of transactional operations that might – depending on the concrete case considered – be qualified as juridical acts. As a drawback for embracing such a vast categorization, the outlines of a smart contract's essence are forcefully blurred and often out of our focus: while some contributions still cling to the proto-industrial prototype of the vending machine (which is not particularly challenging from a legal perspective), others evoke sophisticated, computerized procedure that nevertheless lack the economic architecture representing the ubi consistam of private autonomy.

In this fragmented reality, a brand-new movement seems to emerge: being increasingly identified with codex operating in the blockchain, smart contracts are currently supposed to regulate the economic operations that take place on this platform and to automatically execute all the transactions mandated by their directives. Through this escamotage, the relation between the two technologies is inverted, and the blockchain is set to become the minimum common denominator for any smart contract, which is expected to ontologically operate within the blockchain.²

The inevitable outcome is that any contract can get “smart”, as long as its information transmigrate into the blocks.³

Such a position unveils, however, the major shortcomings of this narrative, and the fragility of its foundations: the legitimacy of the whole smart contract framework is based on the idea that the blockchain will radically redefine and

¹ See, ex multis, E. MIK, Smart Contracts: Terminology, Technical Limitations and Real World Complexity, 9 Law, Innovation & Technology 269 (2017).

² MIK, cit.: “purportedly, smart contracts are contracts that are represented in code and executed by computers. They are not only formed online but their very performance is enabled and *guaranteed* by a network of decentralized, co-operative computer nodes, known as blockchains”. See also L. PAROLA, P. MERATI and G. GAVOTTI, *Blockchain e smart contract: questioni giuridiche aperte*, in Contratti, 2018, 681, 683.

³ On the idea that “contracting has always been about the fine print” and that “code can capture the fine print” see J. S. GANS, The fine print in smart contracts, www.nber.org/papers/w25443, 2019.

What is wrong in the debate about smart contracts

pervade any aspect of individual relationship, originating the true “contract law 2.0”.

All these conjectures are, on the contrary, and against this bedrock, doomed to fail if blockchain will reveal itself – as many are hypothesizing – as the proverbial solution in search for problems: in spite of the huge enthusiasm accompanying the developments of this technology in its early days, blockchain still struggles to find far-reaching applications. If its contribution to the improvement of commercial transactions should turn out to be marginal, its role as a means to revolutionize contract law should be questioned as well.

As a consequence of the current technical and economic uncertainty of the phenomenon, the debate about its legal implication is equally precarious, and defending the idea of upcoming diffusion of these technologies require some sort of a leap of faith. On one side, “true believers” of smart contracts and blockchain magnify these tools, their potentials, and capacity to bring the automation of contract to its limits; they promote the use of technologies that can – allegedly – predict a huge amount of variables to provide highly sophisticated solutions. On the other side, there are scholars unwilling to endorse this technologic miracle: they display doubts on the capacity of smart contracts to encircle all the different facets that characterize traditional contracts – and their bargaining process – suggesting a more cautious approach.

The framework is equally heterogeneous when the implications of smart contracts for social developments are considered. Many contributions stress the macroscopic issues that the widespread deployment of smart contracts might raise in terms of governance and control by public entities and their impact on the democratic structures of power, forestalling the rise of new mechanisms for the resolution of disputes, subtracted from the traditional jurisdiction of public law.

Despite the presence of a significant strand of research investigating the impact of smart contracts in terms of public-private powers interaction, the vast majority of scholarly contribution on the topic is certainly focused on the impact of this technology on private relations. Here, the definitional problem arises once again: it is, in particular, necessary to decide whether a smart contract can be qualified as a true agreement or if (on the contrary) the notion is nothing but a “fancy name” to identify those – already existing – codices that interact with blockchain platforms.

The debate therefore moves to the characterization of “smart contracts” as proper contracts according to private law – or, rather, as innovative protocols¹ to transfer data. If the second alternative is accepted, then the impact of smart contracts on the juridical reality is drastically diminished: the expectation that smart contracts may radically redefine the dynamics of digital exchanges, limiting parties’ autonomy in the execution of their obligations and ultimately nullifying any contractual risk, is confuted at its very own roots. The same fate goes for all claims regarding the capacity of such contracts to overcome the vast majority of problems characterizing “human” negotiations and to promote a high-level of protection for “weak” agents in the market, and in primis for consumers, who, according to “technological evangelists”, might be adequately empowered – by

¹

Roberto Pardolesi and Antonio Davola

these emerging opportunities – to finally counteract the information asymmetry that constitute the traditional source of disparity in B2C contracts. In addition to that, opponents claim that, since smart contracts are incapable of effectively displaying individual preferences, an unintended effect of their consistent adoption would be depriving regulators of significant pieces of information on consumers' priorities and *modus operandi*. Also, further concerns invest the potentially discretionary evaluations that algorithms can operate, and their adequacy to conform to social values such as transparency and substantive justice. Whether the take-off of smart contracts would overall lead to an effective increase of consumer protection within the legal system is not that easy to assess.

To tell the truth, no aspect of this debate seems truly original to us: many of these topics have already been discussed within disputes over the impact of automated decision making technologies on the law.

On the opposite, if the “contractual path” is embraced, then the compatibility of smart contracts with the provisions of contract law is immediately relevant.¹ The issue can be addressed from different points of view, including – but not limited to – the way interpretation rules are supposed to operate for smart contracts; how the adequacy of consent is proven when the agreement is concluded (especially if the contract operates as a long-term one); how liability should be allocated in case of malfunctioning of the smart contract, and which level of burden of proof should be required from claimants and defendants in order to succeed in court.

All these topics, nonetheless, deserve attention only if we believe that a smart contract is, at the very end, a (type of) contract: if, on the contrary, they represent mere tools susceptible to be encompassed within the (traditional) contractual practice – as part of legal scholarship defends –, then all these questions are devoid of their primary foundations.

3 On Virtues and Vices of Innovation

Against the fragmentation of the existing debate – as a result of the disordered overlap and intertwining of heterogeneous contributions – some clarification is therefore opportune. In our view, two profiles should be further investigated: firstly, how the relation between blockchain (i.e. the cryptographic technology) and smart contracts (in its phenomenological meaning) should be assessed. Secondly, smart contracts' very own nature should be analyzed, in order to ascertain whether to qualify them as (sort of) digitalized contracts.

3.1 *Relation between blockchain and smart contracts*

¹ FINOCCHIARO, cit.; PAROLA, MERATI, GAVOTTI. *supra* fn. ...

What is wrong in the debate about smart contracts

Considering the first aspect, we hold true that the mantra of a stringent relation between blockchain and smart contract should be dissipated, in the sense that the latter are not ontologically dependent from the existence of blockchain-like technologies. This position derives from a series of persuasive considerations.

Firstly, the use of binary codes to incorporate and computerize parts of a contract is not a brand-new phenomenon: for example, the usage of electronic format to digitally communicate was already diffused in the product chain before internet and e-commerce got massively exploited through EDI (electronic data interchange) technologies. Then, building on this experience, data-oriented contracts – in which the parties have expressed one or more terms or conditions of their agreement in a manner designed to be processable by a computer system – spread out; and still, the interpretation of the contractual terms and parties' intention has been demanded to the traditional juridical system.

It is indeed true that Szabo's informatic protocols (in their hypothetical functioning) were radically different from the previously existing solutions: they were supposed to circulate on the internet, transposing and digitalizing any aspect of the contractual process, from negotiation to execution. Szabo's digitalized contracts would have operated without any external support, nor human assistance, encompassing any relevant legal aspect: they would have been, truly, supercontracts (if you prefer, bot contracts...).

Until 2008, though, Szabo's conception was little more than a project. Then, finally, the development of the blockchain and the bitcoin revolution offered his followers an open-ended technology, suitable to be virtually applied to any form of "trusted transactions", including contracts.

Our referring to trust is not extemporaneous: in order to properly operate, traditional contracts must be binding for the parties: hence, the traditional recourse of the parties to external systems of remedies, adjudication, etc. Against such a traditional framework, when a contract is digitalized within a decentralized ledger, it acquires immutability and the correct execution of the contractual obligations is safeguarded. In other terms, "once the computers determine that the requisite state has been achieved, they automatically perform data-oriented or computable contracts ... The computers in the blockchain network ensure performance, rather than any appendage of the state. And, because blockchains run on a distributed network of independent nodes, with no central control point, a litigant seeking to enjoin performance of a smart contract has no one to sue".¹

The chain constituted through electronic blocks represents, therefore, the applicative dimension that Szabo's vision was devoid of. In the wake of the enthusiasm generated by the rise of cryptocurrencies, many scholars believed that a so long desired paradigm-shifter was found.² In the blockchain, smart contracts are supposed to be protected from manipulation, to ineluctably execute the obligations they encompass and at the same time avoid all the vagaries that characterize the "forensic lottery": they are, so to say, engineered contracts.

¹ WERBACH and CORNELL, *supra* fn. ..., 332.

² SAVELYEV, *supra* fn. ..., 121.

Roberto Pardolesi and Antonio Davola

The implementation of smart contracts obviously presents potential drawbacks (e.g., if the contract is truly unmodifiable, how should an error be corrected, or how should a lack of information be filled without any possible legal intervention?);¹ these problems, though, are often overlooked, due to the frenzy generated by the speculation on cryptocurrencies. Ultimately, it is undeniable that the introduction of the blockchain brought once again the debate on smart contracts at the center of the legal and regulatory arena.

Yet, this does not mean that the blockchain is a constitutive element of any smart contract, or the sole means to develop this instrument: if, on one side, blockchain offers the possibility to have self-enforcing contracts including transparent and immutable information, the profound uncertainty surrounding the general applicability of decentralized ledgers should not be overlooked as well. When this aspect is considered, it is intuitive to grasp an inconvenient truth: the automatic execution of contractual obligations – when the conditions set by the parties and written in the code are met – is not ontologically derived from the use of the blockchain: it is, rather, the result of the parties' choice to devote an informatic technology to the performance of the (previously concluded) arrangement, after agreeing on its conditions.² What truly counts is that the parties trust the reliability of the automatic system, be it the blockchain or another one.

We might, therefore, assume that the blockchain is not the condition *sine qua non* for the functioning of smart contracts, but just one of the possible tools for their implementation; if smart contracts are meant to spread in the legal practice, this might as well happen through technologies, other than the blockchain,³ that will reveal themselves as more suited to adapt to users' needs.

3.2 *Analysis of smart contracts*

If the relation between blockchain and smart contracts can be disentangled with a minor burden, assessing whether smart contracts should be qualified - *ratione materiae* – as proper contracts is a much more difficult task, given the vast heterogeneity of aspects to take into account.

It goes without a doubt that a smart contract does not operate as a traditional one whenever the automated protocol administers specific fragments or parts of a general, wider, agreement: one step in the contractual iter – as technological and clamoring might be – does not encompass the whole agreement, being rather just a tassel in a more complicated mosaic. When this perspective is considered, the issue is easily solved.

The problem is, though, harsher if we hypothesize that the code is used to gather and regulate the whole agreement (as the result of the parties' private autonomy)

¹ RODRIGUES, *Law and the Blockchain*, *supra* fn. ..., 714 ss

² DI CIOMMO, *supra* fn. ...

³ E.g. in 2017 the CodeX- Stanford Center for Legal Informatics launched a White Paper on the s.c. Legal Specification Protocols. See *Developing a Legal Specification Protocol: Technological Considerations and Requirements*, <https://law.stanford.edu>, Jan 2019 (p. 27 ss.).

What is wrong in the debate about smart contracts

and that the software is able to operate autonomously and to self-sufficiently execute the contract in any respect. In this case, we are inspecting the projection of the contract into a fully computerized process. The plan of the parties is reversed into a program, which is meant to perform all the related procedures without external interventions: the agreement is, in some sense, the trace upon which the code is written, as a crystallization of the parties' 'meeting of the minds' in binary form. This might not be a contract, but it certainly looks like its operative form.

Proceeding even further on this path, we might evaluate the characteristics of those situations, when consent originates since the beginning in code form. As Surden pointed out,¹ no existing provision of contract law prohibits the parties to express their contractual obligations as data-oriented representations (that might as well be included in the blockchain): any economic operator might, for example, structure a code providing that, when (and if) certain conditions are met, some goods are transferred to another individual. Such a phenomenon would not create significant uncertainties for contractual law, but for the fact that, in the example just made, the software is autonomously elaborated by one party, and the counterparty is required to merely provide her approval in order to make the contract binding: through this escamotage, the essential moment of the agreements escapes the clutches of computerization.

This consideration has two consequences: on one side, it invites scholars to address the divide between unilateral and bilateral contracts, and its implication within the realm of smart contracts. On the other side – and more importantly – it unveils that the traditional narrative of smart contracts as tools to reduce time and costs of transactions might be superficial, as soon as the creation of a (true) smart contract requires the contribution of a plurality of people. Someone who wants to launch a contractual initiative needs first, and foremost, a software developer² to transpose the instructions related to the various aspects of the agreement into the virtual architecture, to build a data model (structured according to directives and conditionals) for the offer to operate, considering also different forms of interaction with the counterparties. Hence, the possibility of radically disintermediating the contract is already confuted at its roots. Consider that, moreover, the developer must also be able to relate with another operator, in charge of establishing the conditions for the access to the platform where the code will operate (and where the general public, or the counterparty, will be able to find and eventually conclude it). Lastly, users acting on the platform must be able to interact with the "proposal" if they wish to amend its conditions before concluding the contract, as it traditionally happens during negotiations: this requires the presence of other intermediaries on their side too.³

¹ WERBACH e CORNELL, *supra* fn. ..., 342 ff.

² Note that this entails significant issues in terms of trust as well. See A. WALCH, *In Code(rs) We Trust: Software Developers as Fiduciaries in Public Blockchains*, www.ssrn.com, 2018.

³ A thoughtful investigation of this topic has been offered by S. COHNEY, D. HOFFMA, J. SKLAROFF e D. WISHNICK, *Coin-Operated Capitalism*, www.ssrn.com, 2018, who investigated a series of 50 2017 'Initial Coin Offerings', ICOs, conducted on smart contracts. Despite the offering claim that "code does have the potential to become a substitute and complement for old-fashioned legal governance in financial contracting", the Authors ultimately observed that "potential does not mean reality. Our study shows just how far code falls short of expectations".

Roberto Pardolesi and Antonio Davola

This means first and foremost that we are facing a multi-level process, entailing (at the very least) significant costs for the parties. In order for these costs to be justified, economies of scale are therefore pivotal: as a consequence, it is reasonable to expect that smart contracts might diffuse primarily in “take it or leave it” scenarios where costs are lower, such as in the case of standard forms.¹

Apart from mass transactions, these tools might find space in ‘framework’ contracts as well, where the general obligations of a social group are collectively regulated (with a minor degree of detail): in this field, smart contracts would eventually typify mass models operating as common templates (and maybe as decentralized applications). In these contexts, given that the software is able to automatize the procedures for the execution of the obligation, major attention could be devoted to developing (standardized) technologies able to elaborate and entrust those same obligations to the software.

The obvious consequence is that traditional, individual contracts shy away from this logic: the presence of heterogeneous needs of the parties, looking for the (often unique) equilibrium of their (possibly conflicting) social, economic and geographical constraints, is difficult to conciliate with the exigency of standardization that would make smart contracts economically viable.

The existence of a proper category of smart contracts, as prospected in its most enthusiastic fashion by scholars, reveals then itself as a fascinating exercise of imagination: everything that is (even vaguely) related to code, bit, and contract, is forcefully conducted within the related debate. We believe, though, that this vast amount of human capital (could, and) should embrace a different and more robust perspective.

4 Looking forward: with caution...

Now, the ultimate question arises: should we conclude that the literature on smart contracts is hopelessly misguided, or can we just try to re-frame the problem, in order for our approach to the topic to evolve?

The critical evaluations we operated on the traditional approach to the legal analysis of smart contracts do not aim at marking what has been written until now as inadequate or detached from reality. Our goal is, rather, to provide some observations in order to bring the debate within the scope of a rigorous methodology. This, to avoid that the excitement over a topic – that is undoubtedly fascinating – ends up undermining the fundamentals of a trustworthy analysis, which is always required from legal scholars.

In our view, two basic conceptual misinterpretations currently affect the debate on smart contracts (and similar technologies).

Firstly, we believe that it is fundamental to radically distinguish the analysis of technologic advance from the investigation on the renewal of the contractual

¹ P. D. FILIPPI and A. WRIGHT, *Blockchain and the Law: The Rule of Code*, Harvard University Press, 2018, discuss about “mass-produced smart contracts”.

What is wrong in the debate about smart contracts

framework. As we already underlined, the rise of “supercontracts” - able to autonomously (and automatically) handle all the heterogeneous aspects characterizing contract practice amongst individuals - is unlikely to happen (at least) in the short term. As a consequence, it is trivial to indulge into disruptive reinterpretations of contracts law: it is first and foremost necessary to comprehend the true nature of the technology at stake and then, eventually, to evaluate the opportunity of a shift in the traditional interpretative paradigms of contract law, knowing that this change must be based on the actual developments of these innovations.

As a corollary, it is opportune to overcome the epidemic belief, according to which any technologic innovation shall be followed by a legal change. Despite emerging technologies’ capacity to promote radical changes in the society, legal operators should always keep in mind that these innovations require a modification of the *ius quo utimur* only if they create original forms of interactions, (potentially legal) relationships amongst individuals, and ultimately new risks within the society, that cannot be addressed through the *ius conditum*. Of course, it is equally possible for a technology to qualitatively innovate an existing phenomenon by exposing the shortcomings of the previously existing regulation, and its supervened inadequacy to pursue those values that it is (and it was since its origin) meant to protect. Yet, in order for this to happen, such a technology must, at least, attain a significant degree of relevance (both in diffusion and utilization) in the society: in the case of smart contracts, it seems that this scenario is still distant.

The second conceptual misinterpretation we want to address concerns the constant overlap between policy and legal interventions in the debate over smart contracts. If it is true that legal analysis can be (as it often happens in the debate amongst American scholars) aimed at the provision of policy indications, it is equally proper to acknowledge that some problems introduced by emerging technologies require evaluations of *latu sensu* political opportunity (e.g. the choice regarding the degree of reliability that a technology deserves): in these cases, the legal analysis should follow – and not anticipate – the political choice.

Considering these (still unsolved) issues, a viable solution for legal scholars could be “giving up” on qualifying smart contracts as a general topic, and to rather focus their investigations on the specific effects that the implementation of an innovation (e.g. blockchain) has on the contractual practice. In doing so, though, contract law should be declined according to its traditional canons and categories, and these technologies should be seen for what they currently are, i.e. segments of pre-existing relational models: they can definitely enhance contractual structures in terms of efficiency but are yet unable to modify their ontological characteristics.

Once clarified that there is a substantial difference between technologic innovation and modification of the regulatory framework, it might therefore be preferable – for those who seriously want to engage in Law&Technology – to focus on the implications of each specific phenomenon (in *primis*, cryptocurrencies) for the legal system. In doing so, the temptation of a generalist approach should always be avoided: the constant stretch towards “disruptive” topics cannot endanger the critical and methodological rigor that must animate legal research.

Roberto Pardolesi and Antonio Davola

There will be plenty of time to establish a solid debate on smart contracts in the future, when the time (truly) comes; as of now, there is probably no need to hurry and fall into inaccuracy.

References

1. RODRIGUES, U.S.: Law and the Blockchain. In: *Iowa Law Review*, 104, p. 679 ff. (2018).
2. WEBER, R.H.: "Rose is a rose is a rose is a rose"- what about code and law? In: *Computer Law & Security Review*, p. 701 ff. (2018).
3. RASKIN, M.: The Law and Legality of Smart Contracts. In: *Georgetown Law Technology Review*, 1, p. 305 ff. (2017).
4. GITTI, G.: Robotic Transactional Decisions. In: *Osservatorio del diritto civile e commerciale*, p. 619 ff. (2018).
5. HACKER, P., LIANOS, I., DIMITROPOULOS, G., EICH, E.: *Regulating Blockchain: Techno-Social and Legal Challenges*, Oxford University Press, Oxford (2018).
6. SUSSKIND, R., SUSSKIND, D.: *The Future of the Professions: How Technology will Transform the Work of Human Experts*, Oxford University Press, Oxford (2015).
7. FINOCCHIARO, G.: Il contratto nell'era dell'intelligenza artificiale. In: *Rivista trimestrale di diritto e procedura civile*, p. 441 ff. (2018).
8. FENWICK, M., KAAL, W.A., VERMEULEN, P.M.: Legal Education in a Digital Age. Why 'Coding for Lawyers' Matters. Available at www.ssrn.com (2018).
9. CLACK, C.D., BAKSHI, V.A., BRAINE, L.: Smart Contract Templates: Foundations, Design Landscape and Research Directions. In: *Cornell University Computing Research Repository*. Available at <http://arxiv.org/pdf/1608.00771v2.pdf> (2016).
10. KÖLVART, M., POOLA, M., RULL, A.: Smart Contracts. In: KERIKMÄE, T., RULL, A. (eds.): *The Future of Law and eTechnologies*, Springer, Berlin, p. 133 ff. (2016).
11. MIK, E.: Smart Contracts: Terminology, Technical Limitations and Real World Complexity. In: *Law, Innovation & Technology*, 9, p. 269 ff. (2017).
12. DI CIOMMO, F.: Gli Smart Contract e lo smarrimento del giurista nel mondo che cambia. Il caso dell'High Frequency Trading (HFT) finanziario, forthcoming.
13. FIMMANÒ, F., FALCONE, G., (eds.): *Fintech, Collana Regole e mercati*, Universitas Mercatorum, Napoli, forthcoming (2019).
14. SCHREPEL, T.: Collusion by Blockchain and Smart Contracts. Available at www.ssrn.com (2019).
15. LEVI, S.D., LIPTON, A. B.: An Introduction to Smart Contracts and Their Potential and Inherent Limitations. Available at <https://corpgov.law.harvard.edu> (2018).
16. PAROLA, L., MERATI, P., GAVOTTI, G.: Blockchain e smart contract: questioni giuridiche aperte. In: *Contratti*, p. 681 ff. (2018).
17. GANS, J. S.: The fine print in smart contracts. Available at www.nber.org/papers/w25443 (2019).
18. SAVELYEV, A.: Contract law 2.0: 'Smart' contracts as the beginning of the end of classic contract law, *Information & Communication Technology Law*, 36, p. 116 ff. (2017).
19. HIGGINSONM, M., NADEAU, M.-C., RAIGOPAL, K.: Blockchain's Occam Problem. Available at <https://www.mckinsey.com/industries/financial-services/our-insights> (2019).
20. SURDEN, H.: Computable Contracts. In: *U.C. Davis Law Review*, 46, p. 629 ff. (2012).
21. BUTERIN, V., e al.: A Next Generation Smart Contract and Decentralized Application Platform, *White Paper on GitHub*.
22. DRUCK, J. A.: Smart Contracts Are Neither Smart Nor Contract. In: *Banking & Financial Services Policy Report*, p. 5 ff. (2018).
23. LEVY, K.E.C.: Book-Smart, "Not Street-Smart: Blockchain-Based Smart Contracts and The Social Workings of Law. In: *Engaging Science, Technology, and Society*, 3, p. 1 ff. (2017).
24. THE ECONOMIST: Not-so-clever contracts. Available at www.economist.com/news/business/21702758-time-beieconomist,ng-least-human-judgment-still-better-bet-cold-hearted, 28 Jul. 2016.
25. WRIGHT, A., DE FILIPPI, P.: Decentralized Blockchain Technology and the Rise of Lex Cryptographia. Available at <https://ssrn.com/abstract=2580664> (2015).

What is wrong in the debate about smart contracts

26. REIJERS, W., O'BROLCHAIN, F., HAYNES, P.: Governance in Blockchain Technologies & Social Contract Theories, Ledger, Vol. 1 (2016).
27. SCOTT, B.: Visions of a Techno-Leviathan: The Politics of the Bitcoin Blockchain. Available at <http://www.e-ir.info/2014/06/01>.
28. O'DWYER, R.: The Revolution will (not) be Decentralized: Blockchain. Available at <http://commonstransition.org/the-revolution-will-not-be-decentralised-blockchains/> (2015).
29. DUROVIC, M.: Law and Autonomous Systems Series: How to Resolve Smart Contract Disputes - Smart Arbitration as a Solution. In: Oxford Business Law Blog, available at <https://www.law.ox.ac.uk>, (2018).
30. HACKE, A.: Law and Autonomous Systems Series: Micro-Justice and New Law? "Swarm Arbitration" as a Means of Dispute Resolution in Blockchain-Based Smart Contracts. In: Oxford Business Law Blog, available at <https://www.law.ox.ac.uk>, (2018).
31. GREENSPAN, G.: Beware of the Impossible Smart Contract. Available at <https://www.the-blockchain.com> (2016).
32. ALLEN, T., WIDDISON, R.: Can Computers Make Contracts? In: Harvard Journal of Law & Technology, 9, p. 25 ff. (1996).
33. WERBACH, K., CORNELL, N.: Contracts Ex Machina. In: Duke Law Journal, p. 339 ff. (2017).
34. STARK, J.: Making Sense of Blockchain Smart Contracts. Available at <http://www.coindesk.com/making-sense-smart-contracts> (2016).
35. CUCCURU, P.: Blockchain ed automazione contrattuale. Riflessioni sugli smart contract. In: La nuova giurisprudenza civile commentata, 1, p. 107 ff. (2017).
36. SWANSON, T.: Great Chain Of Numbers: A Guide To Smart Contracts, Smart Property And Trustless Asset Management (2014).
37. PASQUINO, V.: Smart Contracts: caratteristiche, vantaggi e problematiche. In: Diritto e processo, p. 11 ff. (2017).
38. DI SABATO, D.: Gli smart contracts robot che gestiscono il rischio contrattuale. In: Contratto e impresa, p. 328 ff. (2017).
39. HOLDEN, R., MALANI, A.: Can Blockchain Solve the Holdup Problem in Contracts?, University of Chicago Coase-Sandor Institute for Law & Economics Research Paper No. 846 (2017)
40. HOFFMAN, D.H.: Relational Contracts of Adhesion. In: The University of Chicago Law Review, 85, p. 1396 ff. (2018).
41. FAIRFIELD, J.: Smart Contracts, Bitcoin Bots, and Consumer Protection. In: Washington and Lee Law Review Online, 71, p. 2 ff. (2014).
42. GILLESPIE, T.: Can an Algorithm Be Wrong? In: LIMN, 2 (2012), available at <http://limn.it/can-an-algorithm-be-wrong>.
43. SCHOLZ, L.: Algorithmic Contracts. In: Stanford Technology Law Review, 20, p. 128 ff. (2017).
44. DAVIS, K.E.: Contracts As Technology. In: New York University Law Review, 88, p. 83 ff. (2013).
45. CASEY, A., NIBLETT, A.: Self-Driving Laws. In: University of Toronto Law Journal, 66, p. 429 ff. (2016).
46. CHIERICI, M.: Gli Smart Contract: profili giuridici con una struttura informatica. Available at <http://www.salvisjuribus.it> (2018).
47. KOLBER, A.J.: Not-So-Smart Blockchain Contracts and Artificial Responsibility. In: Stanford Technology Law Review, 21, p. 198 ff. (2018).
48. CAGGIANO, I.: Il contratto nel mondo digitale. In: La nuova giurisprudenza civile commentata, p. 1152 ff. (2018).
49. LIPSHAW, J. M.: The Persistence of "Dumb" Contracts. Available at www.ssrn.com (2018).
50. SURDEN, H.: Computable Contracts. In: U.C. Davis Law Review, 46, p. 629 ff. (2012).
51. GIANCASPRO, M.: Is a 'smart contract' really a smart idea? Insights from a legal perspective, Computer Law & Security Review, 33, p. 825 ff. (2017).
52. CodeX - Stanford Center for Legal Informatics: Developing a Legal Specification Protocol: Technological Considerations and Requirements. Available at <https://law.stanford.edu> (2019).
53. WALCH, A.: In Code(rs) We Trust: Software Developers as Fiduciaries in Public Blockchains. Available at www.ssrn.com (2018).
54. COHNEY, S., HOFFMA, D., SKLAROFF, J., WISHNICK, D.: Coin-Operated Capitalism. Available at www.ssrn.com (2018).
55. FILIPPI, P. D., WRIGHT, A.: Blockchain and the Law: The Rule of Code, Harvard University Press, Cambridge (U.S.A.) (2018).

56. CORNELIUS, K. B.: Smart contracts and the Freedom of Contract Doctrine, *Journal of Internet Law*, 5, p. 3 ff. (2018).

Financial Transaction Data for the Nowcasting in Official Statistics

Transazioni elettroniche di pagamento per le previsioni a breve nella Statistica ufficiale

Righi A., Ardizzi G., Gambini A., Iannaccone R., Moauro F., Renzi N. and Zurlo D.

Abstract The paper provides the first results of an assessment of the potentiality of the use of financial transaction data in terms of forecasting power for macroeconomic variables. The data sources are series from the Payment system data electronically recorded through clearing and settlement circuits managed by the Bank of Italy and series from anti-money laundering aggregate UIF reports. The trials focused both the real-time nowcasting of quarterly value added and turnover of services by ADL models and the forecasting of private consumption using a VAR model. Results showed that new series improve the accuracy of estimates but a preliminary accurate intervention analysis phase is needed.

Abstract Il lavoro fornisce i primi risultati per una valutazione della potenzialità di utilizzo delle transazioni finanziarie elettroniche in termini di potere di previsione per le variabili macroeconomiche. Le fonti sono le serie del Sistema di pagamenti registrate elettronicamente attraverso circuiti di compensazione e regolamento della Banca d'Italia e le serie delle segnalazioni anti-riciclaggio aggregate dell'UIF. Si è realizzato sia il nowcasting in tempo reale del valore aggiunto e del fatturato dei servizi trimestrali con modelli ADL; sia la previsione dei consumi privati con un modello VAR. I risultati mostrano che le nuove serie migliorano l'accuratezza delle stime ma è prima necessario realizzare un'accurata fase di intervention analysis.

Key words: Big data, ADL model, VAR model, real-time analysis, intervention analysis

¹

Righi Alessandra, Istat; righi@istat.it
Ardizzi Guerino, Banca d'Italia, guerino.ardizzi@bancaditalia.it
Gambini Alessandro, Banca d'Italia, alessandro.gambini@bancaditalia.it
Iannaccone Roberto, Istat, iannacco@istat.it
Moauro Filippo, Istat; moauro@istat.it
Renzi Nazzareno, UIF- Banca d'Italia, Nazzareno.Renzi@bancaditalia.it
Zurlo Davide, Istat; zurlo@istat.it

Introduction

The use of big data for nowcasting and forecasting macroeconomic indicators has recently attracted the attention of institutions, researchers, stakeholders and policy makers. According to Eurostat the nowcasting of macroeconomic indicators is the field in which big data can play a decisive role in the future (Baldacci et al. 2016; Kapetanios et al. 2017a; 2017b; 2018). Also for the Central banks big data assume a relevant role for the possibility to get new information at short delay with the aim of increasing the forecasts accuracy (Panetta 2018) and there are pieces of evidence that financial transaction data show suitable features to track the short-term evolution of the economic activity (Ardizzi et al. 2018). In Italy, a recent contribution by Aprigliano et al. (2017) addressed for the first time the attention to electronic transactions since their link to traditional economic measures finds solid justifications by the economic theory.

Concerning the use of electronic financial transaction data, a study on the potentiality in term of forecasting power of the new data sources has been launched through a Bank of Italy - Istat collaboration agreement aimed at the production of new series and at studying their possible uses, besides traditional economic series, for macroeconomic aggregates forecasts. New sources concern both the financial transaction data of the exchange circuits and interbank settlement of the Payment system (electronically recorded through clearing and settlement circuits managed by the Bank of Italy) and the anti-money laundering aggregate reports (SARA) that financial intermediaries monthly send to the Bank of Italy Financial Intelligence Unit (UIF).

1 Financial transaction data as a data source

All the actors and the electronic transactions exchanged within the Payment system can be traced back to a scheme in which the transactions that take place in the Market of goods and services are negotiated by banks and intermediation agents and in the post-trading phase, infrastructures and clearing or settling systems operate (Figure 1).

The joint Bank of Italy and Istat project focused on transactions and systems managed in BI-Comp and TARGET2, which are two compensation and settlement circuits, where both credit card payments and giro-payments flow.

In Payment system, there are two main categories of payments: wholesale payments (large-value payments, connected with financial markets flows and refinancing operations with national central banks) and retail payments (mostly related to individual economic activity). As for the retail payments, since 2014 the Single Euro Payments Area (SEPA) has fostered the interoperability among different national clearing and settlement retail systems. The Bank of Italy manages the BI-Comp clearing and settlement system (in compliance with SEPA) which clears the domestic payments on a multilateral net basis. However, payments can be settled

Financial Transaction Data for the Nowcasting in Official Statistics

both in BI-Comp and in TARGET2 (in a retail branch named TARGET2-retail). These data are collected on a daily basis and very timely disseminated by the Bank of Italy on a monthly basis. They concern about 40-50% of total transactions, corresponding to those between customers that do not join the same banking group. The retail non-cash payments settled through BI-Comp and TARGET2-retail add up to about 60% of the total value of retail payments in Italy and 80% considering only electronic payments (Aprigliano et al. 2017). In term of relevance, the total amount of payment flows settled through T2-retail and BI-Comp is about three times the Italian GDP on annual basis.

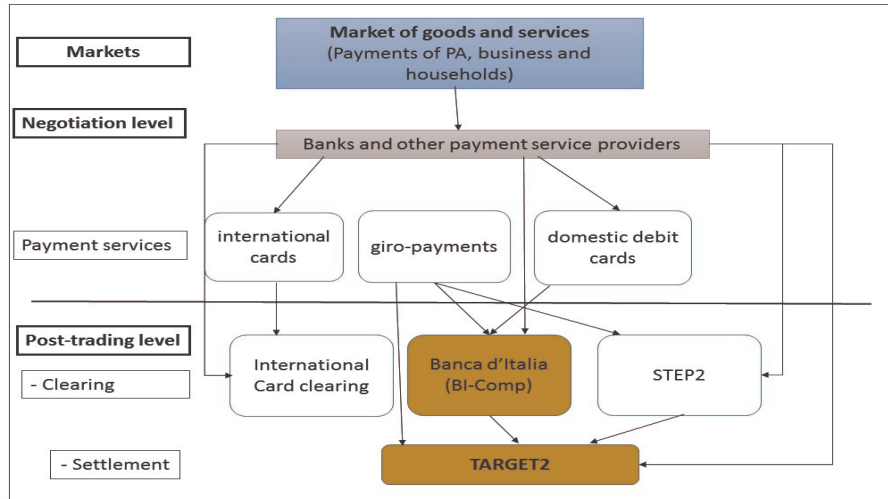


Figure 1: Scheme of the Italian financial transaction data sources and flows in the Market of goods and services of Payment system

The Bank of Italy-Istat collaboration allowed to make a survey of all the flows that can be useful for the forecasting of macroeconomic indicators; this survey resulted in the production of about twenty monthly time series (in amount and in number of transactions) extracted from two inter-exchange and settlement circuits (BI-Comp and TARGET2 retail), broken down by payment instruments (i.e. credit transfers, direct debits, payment cards as POS and ATM, other) and reconstructed backwards since January 2000 to December 2017.

The Payment system data are fully comparable during the considered period (from January 2000 to December 2017) except for some changes in the regulation of the system that create some shocks. The pre-treatment and inspection phase highlighted some periods with particular perturbations for the BI-Comp series; particularly low values are observed since December 2013 for several months, due to the launch of the Pan-European inter-bank exchange and settlement system for private operators (STEP2-T). This new private system, owned by EBA Clearing, initially eroded the representativeness of the BI-Comp series, but this phenomenon has diminished in short term.

The other big data source is the anti-money laundering aggregate reports (SARA) filed by financial intermediaries to the financial intelligence unit (UIF) of the Bank of Italy¹. The database consists of anonymous data relating to all the transactions of an amount equal to or greater than EUR 15,000². They are over 100 million records per year, which are monthly collected (with a delay of around 60 days) and containing information on the municipality where the transaction took place, on the kind of transaction (give/take) and summary information of the customers performing the transaction (e.g. detail in terms of industry of the actors involved in the transactions).

About fifty series, considered for the first time in this context, referring to the total value of the banking transactions, cash operations and domestic or foreign wire transfers are provided for the project. Some series revealed some perturbations in the period from October 2002 to December 2003, due to changes in the technology used for reporting the series of cash/debit amounts. Thus, a preliminary phase of outliers' detections was necessary to decide how to intervene on the breaks of the series.

Elementary data of both the considered data sources were aggregated in compliance with the existing rules General Data Protection Regulation (GDPR) in order to produce time series for our experiments.

2 Impact assessment of the use of financial transaction data on the macroeconomic indicator estimates

The development of impact and quality assessments of the use of transaction data in forecasting was fostered by the participation of the members of the joint Banca d'Italia - Istat collaboration to the project ESSnet Big Data on early estimates of economic indicators. This project was devoted to exploring how a combination of big data sources and existing official statistical data could be used in early estimates (Luomaranta et al 2018). Afterwards, further experiments were conducted and concrete estimates for economic indicators making use of new series are described in the following sections.

2.1 Private consumption

Aimed at forecasting (one-quarter ahead) and nowcasting (current-quarter forecasts) the series of private household consumption for all goods, coming from National Accounts quarterly estimates, we exploited the timeliness of BI-Comp and TARGET2 series. We studied a simple VAR model, actually used in Istat for an

¹ Within this exercise, only reports from banking intermediaries were used.

² In keeping with the Italian anti-money laundering regulation, some transactions below this threshold are also reported.

unpublished forecast, with the aim to find out the payments data providing a most accurate forecast.

We seasonally adjusted and pre-treated the series, as they exhibited a change of level at the beginning of 2014 that we treated using an ad hoc regressor called "the ramp effect". Figure 2 shows the result of the pre-treatment of the BI-Comp Total and TARGET2 payments series (amounts); the higher series is the original one and the other is the adjusted one. Figure 3 shows the result of the seasonal adjustment method performed using Demetra+ on the linearized series.

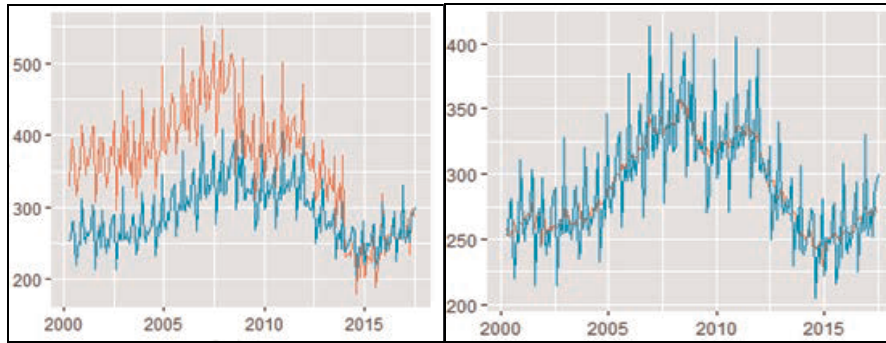


Figure 2: BI-Comp Total and TARGET2 Payments: original and linearized series **Figure 3:** BI-Comp Total and TARGET2 raw and seasonally adjusted data

For the estimates we used VAR models separate for the two components consumption of durable goods (e.g., buying or building a house within the next 12 months, car registration, household purchasing power) and consumption of non-durable goods (e.g., industrial production of consumer goods, the volume of stocks currently held by retail sales companies). The forecasts of the two aggregates are combined to forecast the private consumption growth rate.

We studied the cross-correlation (measured in terms on first differences) between private consumption and transactions and we checked the correct lag to use to introduce these variables in the VAR models. We used the contemporary and delayed series of a lag in the forecasting period in the case of Payment system series, whereas SARA series did not present correlations, except for series relating to transactions in the retail sector. Thus, we added each series, deflated with the GDP deflator, individually to the VAR model for consumption of durable goods and consumption of non-durable goods forecasted (at one and two steps) in pseudo-real-time (period: first quarter 2015 – second quarter 2017).

Results showed that series producing higher gains in terms of accuracy of forecasts are the BI-Comp Total and TARGET2 payments series (gain of +10% in terms of MAE for the forecasting one step ahead) and the series referring to checks. The only SARA series that has determined accuracy gains are the transactions in the retail sector.

2.2 *Early economic indicators*

Since May 1 2018 the quarterly GDP early estimate is officially released in Italy at 30 days by the end of reference quarter and is compiled using the same sources and methods adopted for the full compilation of quarterly National Accounts (released at 60 days). This GDP estimate is obtained by aggregation of more than 50 sub-components of value added. However, as genuine data sources are available only at an annual frequency and after eighteen months, quarterly estimates are obtained by temporal disaggregation and extrapolation methods. A Chow and Lin-type method is adopted (Chow and Lin 1971), which uses short-term indicators as related information and allows depicting the quarterly pattern of the set of quarterly National Accounts (Istat 2008). This approach guarantees high standards to the GDP estimation, since early estimates show low revisions with respect to full-informative releases, even if several indicators are either totally or partially missing at the T+30 round. Nevertheless, concerning the service sector, no indicator is sufficiently timely for early estimates. Thus, the availability of financial transaction series is particularly relevant to set forecasting (one-step ahead) exercise of the value added of services.

In the presented exercise we made reference to autoregressive distributed lag models of order ADL(1,1) conducted at a quarterly frequency. The one-step ahead forecasts refer to 16 quarters over the period from the first quarter of 2012 to the fourth quarter of 2015, as well as forecast errors are evaluated with respect to the T+60 releases based on information of the full set of indicators. For a comparative evaluation, an exercise of pure extrapolation on data of service value added has also been carried out through a first-order autoregressive model AR(1). This exercise provides first evidence on benchmark values to which refer to in the evaluations of the new indicators in terms of mean absolute errors (MAE). Results show that some series reduce the mean error statistics of the AR(1) benchmarks, even if only after a careful intervention analysis for outlier removal. The reduction ranges around 12% relatively to both quarterly and annual growth rates. This gain is in line with that obtained using traditional short-term indicators such as the industrial production index.

We developed other exercises with respect to the turnover of the Services. The recursive estimates conducted with the Payment system's series (in the period from the third quarter of 2011 to the second quarter of 2017) produce lower mean forecast errors than the AR benchmark (1) with gains for the BI-Comp series of 7% in terms of mean errors. The series of collections (+9.9%), TARGET2 (+9.6%) and T2 total customer (+12.8%) resulted particularly performing. The introduction of 40 series of the SARA archive (of sign give and take) from January 2001 to June 2017, distinguished by four types of payment and four sectors of economic activity, also provided positive results. All series are quite performing, with mean absolute errors lower than the benchmark. The series of foreign transfers relating to retail trade showed a gain of 23.9%, a value not far from the gain determined by the industrial production index (31.8 %). The gains showed by the series of total operations relating to retail trade (of sign giving) (+18.4%) are also significant.

3 Discussion and conclusion

The preliminary results of this first systematic evaluation on the use of financial transaction data for the improvement of forecasting models or early estimates in official statistics are positive, even if an accurate intervention analysis for the treatment of the outliers is needed before financial transaction data could effectively be used in forecasting activities. Fluctuations in the series are often due to the changes in regulations and laws ruling the sector of the financial transaction occurred in latest years and they are larger than those observable in commonly used short-term economic indicators regularly used for official early estimates.

Future works, besides the extension of the test-series to other economic variables and the design of a new modelling strategy, concern the attempt for expanding the database on transaction data managed by the central bank. Negotiations in this direction with private operators of the Payment system are underway.

References

1. Aprigliano V., Ardizzi G., Monteforte L.: Using the payment system data to forecast the Italian GDP, Banca d'Italia, Temi di discussione (Working papers), 1098 – February (2017)
2. Ardizzi G., Emiliozzi S., Monteforte L., Marcucci J.: News and consumer card payments, Banca d'Italia, Working Papers series (forthcoming)
3. Baldacci E., Buono D., Kapetanios G., Krische S., Marcellino M., Mazzi G. L., Papailias F.: Big data and macroeconomic nowcasting: from data access to modelling, Eurostat statistical books (2016)
4. Chow, G.C., Lin A.: Best linear unbiased interpolation, distribution and extrapolation of time series by related series. *Rev.Econ.Stat.* 53, 372–5 (1971)
5. Istat: Quarterly national accounts inventory, sources and methods of Italian quarterly accounts (2008), available at <https://ec.europa.eu/eurostat/documents/24987/4253464/IT-QNA-Inventory-ESA95.pdf/5f7d98d8-2734-448c-9c5f-75845e648bc1>
6. Kapetanios G., Marcellino M., Papailias F.: Big data conversion techniques including their main features and characteristics, Eurostat statistical working papers, July (2017)
7. Kapetanios G., Marcellino M., Papailias F.: Filtering techniques for big data based uncertainty indexes, Eurostat statistical working papers, November (2017)
8. Kapetanios G., Marcellino M., Petrova K.: Analysis of the most recent modelling techniques for big data with particular attention to Bayesian ones, Eurostat statistical working papers (2018)
9. Luomaranta H., Puts M., Grygiel G., Righi A., Campos P., Grahonja Č., Špeh T.: Deliverable 6.6: Report about the impact of one (or more) big data source (and other) sources on economic indicators, Work Package 6 - Early estimates of economic indicators, ESSnet Big Data specific grant agreement 2 (SGA-2), pp.36-42 (2018)
10. Panetta F.: Big data and machine learning technology for central banking, speech at the Conference Harnessing big data & machine learning technology for central banking, Roma, March (2018)

On the examination of a criticality measure for a complex system in a forecasting perspective

Esame di una misura di criticità per un sistema complesso in una prospettiva previsiva

Renata Rotondi and Elisa Varini

Abstract Many complex phenomena concerning extreme events causing natural disasters exhibit power-law behavior, reflecting a hierarchical or fractal structure. These phenomena seem to be susceptible to description using approaches inspired by statistical mechanics, particularly approaches involving generalizations of the classical concept of entropy. In this framework we consider the nonextensive Tsallis entropy; by its constrained optimization we derive the expression of the q -exponential distribution which controls the behavior of variables that, in geophysics, can be for instance seismic moment, length of faults, and inter-event times. We estimate the model parameters in the Bayesian perspective and apply this approach to the identification of different dynamical regimes in seismic sequences.

Abstract Molti fenomeni complessi riguardanti eventi estremi causa di disastri naturali si comportano secondo leggi potenza, che riflettono una struttura gerarchica o frattale. Questi fenomeni sembrano poter essere descritti attraverso approcci ispirati alla meccanica statistica, in particolare approcci che coinvolgono generalizzazioni del concetto classico di entropia. In questo quadro consideriamo l'entropia non-estensiva di Tsallis; mediante la sua ottimizzazione vincolata deriviamo l'espressione della distribuzione q -esponenziale che controlla il comportamento di variabili che, in geofisica, possono essere ad esempio il momento sismico, la lunghezza di faglie e il tempo tra eventi. Stimiamo i parametri del modello nell'ottica bayesiana e applichiamo questo approccio all'identificazione di diversi regimi dinamici in sequenze sismiche.

Key words: Tsallis entropy, q -exponential distribution, seismic forecast, Bayesian inference

Renata Rotondi

Istituto di Matematica Applicata e Tecnologie Informatiche - CNR, Via Bassini 15, Milano (I) e-mail: renata.rotondi@mi.imati.cnr.it

Elisa Varini

Istituto di Matematica Applicata e Tecnologie Informatiche - CNR, Via Bassini 15, Milano (I) e-mail: elisa.varini@mi.imati.cnr.it

1 Introduction

Entropy is a measure of the disorder present in a physical system, where the notion of disorder is related to the number of configurations (microstates) that the system can take. In the classical statistical mechanics the definition of entropy is due to Boltzmann-Gibbs (BG); the BG entropy S_{BG} is nonnegative, concave and extensive, that is, it has properties that can be expected to be reasonable approximations for systems containing many particles with short range interactions. In the last decades it has become clear that physical entropies different from the BG one are more appropriate for systems with multifractal or hierarchical structure involving long-range forces. In 1988 Tsallis [9] proposed a generalization of the BG entropy which is characterized by nonextensivity, that is, the entropy of the sum of two *independent* sets can be different from the sum of their entropies. This entropy seems to describe quite well a large number of natural and artificial systems even if, as S_{BG} , it cannot be derived rigorously because the complete theory which describes the elementary processes, like the molecular mechanical theory, is still missing. The Tsallis entropy S_q has received an enormous amount of applications in a variety of disciplines including physics, chemistry, geophysics, biology, medicine, economics, informatics, engineering, linguistics and others [10]. In Section 2 we introduce definition and properties of the BG and Tsallis entropy, and we derive the probability distribution, denoted as q -exponential distribution, which maximizes the Tsallis entropy under some constraints about the normalization and the q -expectation value. In Section 3 this distribution is then applied to model the seismic energy, expressed through the magnitude, released by a sequence of earthquakes; its parameters are estimated following the Bayesian paradigm. In particular, the time evolution of the q -index is analyzed to describe the dynamics of some seismic time series and to test its skill in forecasting the strong shocks (Section 4).

2 Non-extensive entropy

Entropy is a fundamental physical and statistical quantity that characterizes the complexity of the dynamic nature of a physical process. Its definition varies according to the discipline - thermodynamics, information theory, astrophysics, probability - in which it is investigated. It can be considered as a measure of our ignorance about the state of the system, or a measure of the level of organization and of how the energy spreads out in the system under investigation: the higher the entropy the more disordered and random the system and the less the energy is concentrated. Let us consider a system that can have K states, respectively with probability p_i , $i = 1, \dots, K$; the Boltzmann-Gibbs (or Shannon) entropy is given by:

$$S_{BG} = - \sum_{i=1}^K p_i \log p_i,$$

Title Suppressed Due to Excessive Length

where $\sum_{i=1}^K p_i = 1$. It can be easily shown that the entropy S_{BG} is nonnegative, concave and extensive, where extensive means that if A and B are two *independent* systems ($p_{ij}^{A+B} = p_i^A p_j^B$), then we have $S_{BG}(A+B) = S_{BG}(A) + S_{BG}(B)$. This entropy is useful for studying systems whose elements interact locally in space and time, but other entropies have to be proposed to investigate systems whose elements interact globally. In 1988 Tsallis proposed the only generalization of S_{BG} that maintains the basic properties but allowing, for $q \neq 1$, nonextensivity. Given:

$$S_q = \frac{1 - \sum_{i=1}^K p_i^q}{q-1},$$

or

$$S_q = \frac{1 - \int_X f(X)^q dX}{q-1}$$

in the continuous case, with $q \in \mathbb{R}$, we have:

$$S_q(A+B) = S_q(A) + S_q(B) + (1-q)S_q(A)S_q(B)$$

for two A and B *independent* systems. It follows that $q = 1$, $q < 1$, and $q > 1$ respectively correspond to the *extensive*, *superextensive* and *subextensive* cases; we note that, for $q \rightarrow 1$, S_{BG} recovers S_q , where q is called “entropic index”.

For the maximum entropy principle the probability distribution which best represents the current state of knowledge, given by the observed data themselves, is the one that maximizes the entropy. Hence we derive the density $f(\cdot)$ that solves the optimization problem:

$$\begin{aligned} \max_f \quad & \frac{1 - \int_X f(X)^q dX}{q-1} \\ \text{subject to} \quad & \int_0^{+\infty} f(X) dX = 1 \\ & \int_0^{+\infty} X f_q(X) dX = \bar{X} \end{aligned} \quad (1)$$

where \bar{X} is called q -expectation (1) and $f_q(X)$ is the escort probability given as follows:

$$f_q(X) = \frac{f^q(X)}{\int_0^{+\infty} f^q(X) dX}.$$

Using the standard technique of λ_1 and λ_2 Lagrange multipliers, the previous problem results in maximizing the following functional:

$$\phi(f, \lambda_1, \lambda_2) = S_q + \lambda_1 \left(\int_0^{+\infty} f(X) dX - 1 \right) + \lambda_2 \left(\int_0^{+\infty} X f_q(X) dX - \bar{X} \right) \quad (2)$$

By imposing $\partial\phi/\partial f = 0$ we obtain, for $1 < q < 2$:

$$f(X) = \frac{1}{\bar{X}} \left[1 - \frac{(1-q)}{(2-q)} \frac{X}{\bar{X}} \right]^{1/(1-q)} \quad X \in \mathbb{R}^+ \quad (3)$$

and for $0 < q < 1$:

$$f(X) = \frac{1}{Z_q} \left[1 - \frac{(1-q)\lambda_1}{c_q} (X - \bar{X}) \right]^{1/(1-q)} = \frac{1}{Z'_q} \left[1 - (1-q)\lambda' X \right]^{1/(1-q)} \quad X \in (0, x_{max})$$

where Z_q, Z'_q are normalizing constants, $c_q = \int_0^{x_{max}} f^q(X) dX$, x_{max} is the cutoff to be imposed whenever the argument of the q -exponential function becomes negative, and

$$\lambda' = \frac{\lambda_1}{c_q + (1-q)\lambda_1\bar{X}}.$$

3 Application in geophysics

Today it is widely accepted that most earthquakes are originated by relative motion of fault planes, whereas different perceptions exist on what controls the energy release. In 2004 Sotolongo-Costa and Posadas [6] proposed a model for earthquake dynamics based on fragment-asperity interaction. It assumes that the space between the two rock blocks that slip one with the other during an earthquake is filled with the residues of the breakage of the tectonic plates that generated that fault; the relative motion of the two blocks can be hindered by the overlapping of two irregularities on fault profiles and by the relative position of the fragments. Consequently, the energy ϵ released during an earthquake scales with the size σ of these fragments: in particular, Sotolongo-Costa and Posadas [6] first assumed that the energy scales with the surface of the fragments, $\epsilon \sim r^2$, whereas later Silva *et al.* [4] chose the volume, $\epsilon \sim r^3$.

In this framework, the nonextensive statistics can be suited to describe the distribution of the surface $X \sim r^2$ of the fragments; replacing X with $(\epsilon/a)^{2/3}$ in (3) we obtain the probability distribution of the energy:

$$f(\epsilon) = \frac{c_1 \epsilon^{-1/3}}{\left[1 + c_2 \epsilon^{2/3} \right]^{1/(q-1)}}$$

where $c_1 = \frac{2}{3} \frac{1}{a^{2/3}} \frac{1}{\bar{X}}$, $c_2 = -\frac{(1-q)}{(2-q)} \frac{1}{\bar{X} a^{2/3}}$, and the seismic energy per unit volume a ($\sim 10^{2.6 \pm 1.86}$) is the proportionality constant between ϵ and r^3 . Finally, if we consider the relationships between seismic energy and moment magnitude M_w or duration magnitude M_D respectively:

$$\log_{10} \epsilon \sim 1.5 M_w \quad \text{and} \quad \log_{10} \epsilon \sim 1.94 M_D,$$

we obtain:

$$f(M) = \log_e 10 \frac{10^M}{\beta} \left[1 - \frac{(1-q)}{(2-q)} \frac{10^M}{\beta} \right]^{1/(1-q)} \quad (4)$$

being $\beta = a^{2/3} \bar{X}$, and $M = M_w$ or $M = 1.3M_D$.

3.1 Bayesian estimation

In the literature there are studies which analyze the magnitude distribution from the nonextensive viewpoint [8]; in general the parameters are estimated through the maximum likelihood method which, in this case, requires numerical methods to solve the system of the first partial derivatives set equal to zero.

On the contrary we follow the Bayesian paradigm and we consider the two parameters β and q as random variables; for computational reasons we reparametrize the distribution (4) by setting

$$\theta = \frac{2-q}{q-1}$$

so that $\theta \in (0, +\infty)$ when $q \in (1, 2)$. The knowledge drawn from the literature is translated into the prior distributions for the parameters: both LogNormal distributions with hyperparameters such that $E_0(\beta) = 25$, $var(\beta) = 64$ and $E_0(\theta) = 2$, $var(\theta) = 4$. The posterior distributions are obtained as the equilibrium distributions of Markov chains generated by the Metropolis-Hastings algorithm with LogNormal proposal distributions depending on the current value of the chain; for instance, for θ parameter at the j -th iteration of the Markov chain, the candidate value $\tilde{\theta}$ is drawn from a LogNormal distribution $p(\theta|\theta_{j-1})$ with mean equal to θ_{j-1} and variance equal to $\kappa_\theta \theta_{j-1}$ where κ_θ is chosen so that the acceptance rate is about 30%. Analogously we carried out for β parameter. In particular, in the case studies described in Sec. 4, we used $\kappa_\beta \in (1.0, 1.3)$ and $\kappa_\theta \in (0.4, 0.6)$. Moreover, some diagnostic tests were performed to check the convergence of each Markov chain to its target distribution, which are implemented in *BOA* R package [5].

4 Case studies

We examine two cases concerning the seismic activity recorded during the years that preceded and followed the most recent destructive earthquakes occurred in Italy: Amatrice-Norcia 2016 and L'Aquila 2009 earthquakes. Our aim is to verify the skill of the q entropic index to identify changes or variations in the physical process of earthquake generation that can be associated with a phase of incoming strong activation.

4.1 Central Italy sequence

The central Italy region, included between the longitude (12.7, 13.5), and latitude (42.3, 43.2) values, was hit in 2016 by a long seismic sequence that carried on in the following two years, and that characterized the seismic activity of the entire Italian peninsula. The main events were: on August 24, at 1:37 AM of M_w 6, and at 2:36 AM of M_w 5.3, on October 26, at 5:10 PM of M_w 5.4 and at 7:19 PM of M_w 5.9, on October 30 of M_w 6.5. More than 13,200 events are recorded into the ISIDE (Italian Seismological Instrumental and parametric Data-base) catalog for the period 1 January 2014 through 30 June 2018; most of the magnitude values are expressed in M_L , the remaining in M_w and only few events are given in M_D . To homogenize the data set we applied orthogonal regressions [2] in order to convert all of magnitudes in M_w ; then we selected the 10,457 events of magnitude $M_w \geq 2$ to guarantee the completeness of the data set. Moreover we partitioned the data into subsets: one for each of the years 2014 (N=125 events), 2015 (N=95), 2017 (N=2448), 2018 (from January to June, N=401), whereas the year 2016 was divided into three periods: the first from January to August 24, before the shock of M_w 6 (N=63), the second from August 24 to October 30, before the shock of M_w 6.5 (N=2803), and the third from October 30 to the end of the year (N=4522 events). It can be shown that in the first months of 2016 the seismic rate was on average similar to the one of the previous years (about 8-10 events per month).

First we performed nonparametric tests to evaluate whether the q -exponential (4.1) and classical exponential distributions estimated are significantly different; that is, we tested samples of fixed size simulated, by the inverse method, from the two distributions under the null hypothesis that the two independent samples were selected from populations having the same continuous distribution (or distributions with equal median). To this end we used signed rank tests like the Mann–Whitney U test or the Kruskal–Wallis test; the tests agree on rejecting H_0 (significant differences) in the years 2016 and 2017 of higher activity.

As first step of the analysis we estimated the posterior mean of q parameter for each subsets; the results, reported in Table 4.1, show that q was slightly decreasing for a long time, and grew immediately after the first shock. A more detailed analysis is required if we hope to improve the forecast skill of the method. At this end we have estimated the q parameter at each event over moving (partly overlapping) time windows of 50 events; the obtained values are represented in Fig. 4.2 (right side).

	2014	2015	2016-pre	2016-intra	2016-post	2017	2018
\bar{q}	1.3974	1.3713	1.3718	1.4857	1.4763	1.4549	1.4414

Table 1 Posterior mean of the q entropic index for different subsets of events.

4.2 L'Aquila sequence

On April 6, 2009 L'Aquila area (Abruzzi) was hit by a strong earthquake of M_w 6.3 that caused around 300 casualties; retrospective studies [1] showed that the sequence had started about five months before the mainshock and continued with more than ten thousand aftershocks up to 2010. More than 4600 events are recorded into the ISIDE catalog, occurred from April 16, 2005 to July 1, 2009, in the area included between the longitude (12.8, 13.8) and latitude (41.8, 43.0) values. Also in this case we transformed all of the magnitudes in M_w and just considered the 4509 events of $M_w \geq 1.8$ to guarantee the completeness of the data set. The seismic rate is similar to that observed in central Italy case (about 10-13 events per month) in the years from 2005 to 2009, but, on the contrary, the size level of seismicity is higher; in fact, it is sufficient to look at Fig. 4.2 (a) to note the large number of events with $3 < M_w < 4$ or $4 \leq M_w < 5$ with respect to the ones in Fig. 4.2 (b).

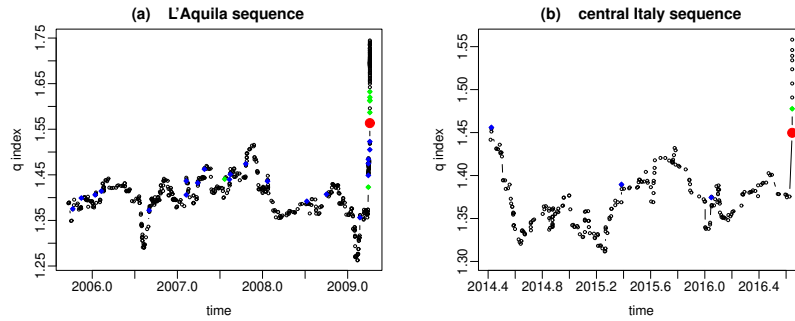


Fig. 1 Estimated values of the q entropic index for L'Aquila sequence (left) and central Italy sequence (right). Colors denote different values of the magnitude of the last event in the moving time window: $M_w \geq 6$ red, $4 \leq M_w < 5$ green, $3 < M_w < 4$ blue.

5 Remarks

In the present article, for sake of space, we have reported only the partial results concerning the analysis of the initial part of the two sequences; nevertheless, some conclusions can be already drawn:

- in both cases q index decreases significantly before the main shock and, in general, a decrease, even if less evident, is almost always present before a medium-high earthquake is recorded;
- in case of L'Aquila a sharp and continuous increase in the q value appears with the occurrence of the March 30, 2009 event of M_w 4; the same does not happen in case of Amatrice earthquake.

Further considerations arise from the detailed analysis of the entire data sets from the viewpoint of the variations of q value; the results are promising but indicate that deeper examination is necessary to check, for instance, their sensitivity to the choices regarding the width of the time windows or the possibility of adopting increasing time windows.

References

1. De Sanctis, A., Cianchini, G., Favali, P., Beranzoli, L., Boschi, E.: The Gutenberg-Richter law and entropy of earthquakes: two case studies in central Italy. *Bull. Seism. Soc. Am.* **101** 3, 1386–1395 (2011)
2. Gasperini, P., Lolli, B., Vannucci, G.: Empirical calibration of local magnitude data sets versus moment magnitude in Italy. *Bull. Seis. Soc. Am.* **103** 4, 2227–2246, doi: 10.1785/012012-356 (2013)
3. ISIDe Working Group: Italian Seismological Instrumental and parametric DatabasE. Istituto Nazionale di Geofisica e Vulcanologia, <http://iside.rm.ingv.it> (last accessed March 2019)
4. Silva, R., França, G.S., Vilar, C.S., Alcaniz, J.S.: Nonextensive models for earthquakes, *Phys. Rev. E* **73**, 026102 (2006)
5. Smith, B.: boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference *J. Stat. Software* **21** 11 doi:10.18637/jss.v021.i11 (2007)
6. Sotolongo-Costa, O., Posadas, A.: Fragment-asperity interaction model for earthquakes. *Phys. Rev. Lett.* **92**, 048501 (2004)
7. Telesca, L.: A non-extensive approach in investigating the seismicity of L'Aquila area (central Italy), struck by the 6 April 2009 earthquake ($M_L = 5.8$). *Terra Nova* **22** 87–93 (2010)
8. Telesca, L.: Maximum likelihood estimation of the nonextensive parameters of the earthquake cumulative magnitude distribution. *Bull. Seis. Soc. Am.* **102**, 2, 886–891 (2012)
9. Tsallis, C.: Possible generalization of Boltzmann-Gibbs statistics, *J. Stat. Phys.* **52**, 479–487, doi: 10.1007/BF01016429 (1988)
10. Tsallis, C., Baldovin, F., Cerbino, R., Pierobon, P.: Introduction to Nonextensive Statistical Mechanics and Thermodynamics, In: Proceedings of the 1953–2003 Jubilee “Enrico fermi” International Summer School of Physics *The Physics of Complex Systems: New Advances & Perspectives*, 1–11 July 2003, Varenna (I)

Knowledge discovery for dynamic textual data: temporal patterns of topics and word clusters in corpora of scientific literature

Estrazione della conoscenza da dati testuali dinamici: evoluzione temporale di argomenti e gruppi di parole in corpora di letteratura scientifica

Stefano Sbalchiero, Matilde Trevisani and Arjuna Tuzzi

Abstract The study aims at comparing two methods for tracing the temporal evolution of topics and keywords in corpora of scientific literature: the well-known Latent Dirichelet Allocation and a new knowledge-based system that has been developed in a functional data analysis unsupervised perspective. Object of the study is a corpus of abstracts of articles published by the American Journal of Sociology over a century (1921-2018). Our study advocates that the two methods might not be seen as alternative but rather as integrable means to improve the interpretation of findings.

Abstract Lo studio mira a confrontare due metodi per tracciare l'evoluzione temporale di argomenti e parole chiave in corpora di letteratura scientifica: la ben nota Latent Dirichelet Allocation e un nuovo sistema basato sulla conoscenza sviluppato in una prospettiva non supervisionata di analisi dei dati funzionali. Oggetto dello studio è un corpus di abstracts di articoli pubblicati dall'American Journal of Sociology nel corso di un secolo (1921-2018). Lo studio propone di considerare i due metodi non come alternativi ma piuttosto come strumenti integrabili per migliorare l'interpretazione dei risultati.

Key words: topic detection, trend analysis, curve clustering, clustering consensus

1 Introduction

Topic modelling and related methods to retrieve relevant topics from texts represent established means to provide an effective content analysis of large corpora nowadays. Methods for topic detection are popular and many articles use Latent Dirichlet

Stefano Sbalchiero, Arjuna Tuzzi
University of Padova, Via Cesarotti 10/12, Padova 35123, Italy e-mail: ste.sbalchiero@gmail.com
e-mail: arjuna.tuzzi@unipd.it

Matilde Trevisani
University of Trieste, Via Tigor 22, Trieste 34124, Italy e-mail: matilde.trevisani@deams.units.it

Allocation (LDA), however, they often merely report short lists of words and briefly describe the content of each topic in order to label them.

When corpora include texts arranged in a chronological order, a crucial issue consists of analysing the evolution of topics through time. In chronological corpora it makes sense to observe the temporal evolution of topics in terms of their probability of occurring and, particularly, highlight topics that experience increasing and decreasing trends (see LDA procedures to identify hot and cold topics in [8]). Incidentally, a topic is composed of a set of words though the temporal evolution of words taken individually does not necessarily reflect the temporal evolution of the topic as a whole.

In previous studies [1, 2, 3, 4] we introduced a statistical learning process within a knowledge-based system (KBS) which, by reconstructing words' life-cycles (time trajectories of word frequencies are interpreted as functional data) and clustering words with similar life-cycles (by a distance-based approach to curve clustering), detects any relevant macro-dynamics underlying word micro-histories.

Differences between LDA and our KBS appear evident: a model-based approach aimed at unveiling topics versus an unsupervised learning procedure aimed at tracing word life-cycles. LDA provides clusters of words that should reflect a topic as they co-occur in texts and that, as a whole, can show an interesting dynamics over time—but that, individually, can evolve differently (even significantly) from the average; on the contrary, KBS provides clusters of words that show a similar trajectory over time—but that do not co-occur in texts and might be ascribed to different topics.

In this study we combine the results of LDA and KBS, with the aim of achieving a better understanding of the effectiveness of LDA for temporal pattern detection. In general, LDA shows to be scarcely informative for tracing topic temporal patterns, leaving aside a rough identification of hot and cold topics. However, LDA might be helpful in interpreting the topic content of historical phases by a transversal reading of topics across the well defined macro-dynamics discovered by KBS.

2 The corpus: American Journal of Sociology

The *American Journal of Sociology* (AJS), established in 1895 as the first U.S. scholarly journal in its field, is considered America's preeminent journal for sociologists from all over the world [5]. The corpus used for this study consists of 4,056 abstracts collected from Volume No. 27, Issue No. 1 (1921) to Volume No. 124, Issue No. 2 (2018), that is, since the first abstracts are available. We excluded material that did not provide information about the journal content or do not belong to scientific papers (e.g., editorials, master heads, errata, etc.).

In pre-processing, the corpus was tokenised and then normalised (by converting uppercase to lowercase). Punctuation marks, numbers and grammatical words (articles, conjunctions, prepositions, pronouns) were removed. In the end, the corpus consists of 25,328 word-types (different words or entries of the vocabulary) and 521,585 word-tokens (occurrences).

3 Topic detection

In a first phase we applied a topic detection procedure [6] to explore the main topics that have appeared in almost a century of AJS publications. In this perspective, a topic is a cluster of co-occurring words and an abstract is a mixture of topics [7]. After applying topic detection, we investigated the main chronological shifts [8] that have occurred in the journal.

Before running LDA, we discarded words with term frequency-inverse document frequency lower than a certain threshold (here, 0.09) to select the most discriminant words and reduce redundancy of highly frequent words as suggested by [9]. LDA analysis was implemented by means of the `topicmodels` R package [9]. To identify the optimal number of topics we used the log-likelihood variation [8] for 2 to 50 topics. The results suggest that it is around 30 (Tab. 1).

Table 1 Excerpt of 5 most probable words for each topic (decreasing order of probability)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
institutional	law	rates	party	students	status	city	inequality
identity	legal	violence	leisure	school	community	cities	income
knowledge	immigrants	decision	global	college	communities	areas	policy
issues	making	crime	local	morale	members	urban	genetic
strategies	justice	making	movement	schools	education	mobility	policies
Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16
movement	race	black	class	protest	family	attitude	women
inventions	negro	white	middle	war	cent	method	men
articles	racial	blacks	market	police	farm	scale	gender
hypotheses	white	racial	classes	movement	families	items	job
science	migration	whites	mobility	mobilization	international	adjustment	employment
Topic 17	Topic 18	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23	Topic 24
organizations	labor	china	prestige	rate	weber	children	religious
organizational	Industrial	conflict	education	fertility	french	family	ethnic
diffusion	industry	chinese	expectations	marriage	attitudes	child	jews
elites	employment	conservative	persons	age	news	families	religion
models	growth	action	scientists	sex	values	parents	catholic
Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30		
professional	sociology	behavior	power	variables	culture		
health	science	personality	network	models	status		
crime	scientific	person	networks	methods	interaction		
government	method	emotional	exchange	variable	marriage		
medical	psychology	conformity	ties	variable	marriage		

Consistent with the idea that topics show different trends through time, we identified the temporal evolution of topics by means of a linear trend analysis [8, p. 5233-5234]. Considering only those with a statistically significant upward or downward linear trend ($p\text{-value} < 1e-4$), we could identify 10 hot/cold topics (Fig. 4).

Although the main results overlap those of a previous research and have already been discussed elsewhere [10, 5], it is worth mentioning that the groups of coldest topics correspond mainly to the methodological development of sociological perspectives and to some specific objects of research whose trajectory shows a peak-like behaviour only in the early period of the journal and then decreased. These topics were very popular in about 20s and 50s (topic 26: the institutionalization process of Sociology as a scientific discipline; topic 15: the measurement of social phenomena; topic 14: the migration between cities and farms; topic 25: the scientific social reflection about health; topic 27: the development of psychological sociology during the 50s). The group of hottest topics is related to articles focussing on a wide range of empirical case studies which highlight the most significant changes that have occurred since the mid-60s (topic 16: the gender revolution; topic 17: the study of organizations; topic 8: inequality from a different perspective as income and related policies; topic 28: mobilization, power and élite; topic 1: the measurement of expertise and institutional knowledge especially until the 70s). The hot topics indicate the birth and recent developments of some sociological topics that clearly indicate how Sociology and sociologists have reacted to new social problems leading to an increase in new sub-disciplines and thus to new specialisations.

4 Word micro-histories versus topic patterns

The KBS first reconstructs words life cycles and second, by clustering words with similar life cycles, detects any exemplary temporal patterns representing the latent dynamics of word micro-histories [4, Par. 9.1.1][3, Sect. 3][2, p. 1579-1580]. In this study, we have chosen a (χ^2 type) double normalisation of word frequencies in order to adjust the uneven document dimension across time as well as to remedy the great disparity in word popularity. Optimal smoothing is achieved with spline order 3 and smoothing parameter 100 ($df = 5.26$) under a roughness penalty of order 1. By the pooled validation approach, curve clustering produces the best set of candidates to cluster number k : 5, 4, 17, 15, 8 (in the order). From the analysis of partitions selected for the candidates, it emerges that the more refined groupings are somehow nested in the coarser ones. We then decided to scrutinize the partitions selected for $k = 5, 15$ (Fig. 1) as they represent a more synthetic vision of the history of the field and, respectively, a more detailed study of the dynamics comparable to the findings of topic detection (Sect. 3). An overview of groups, which have been ordered according to the chronological sequence of temporal patterns, that is, from the cluster of words that have tended to disappear to the cluster of emerging words in the period 1921-2018, is illustrated in Figures 2 and 3. Five fundamental patterns are discovered in the roughest partitioning: 'A', sharply decreasing from 1935 and almost vanishing from 1970; 'B', slowly decreasing after about 1935; 'C', culminant period is '65-'85 after which slowly decreasing; 'D', slowly increasing from about 1965; 'E', sharply increasing from 1980 and still emergent nowadays (Fig. 2). The finest partitioning

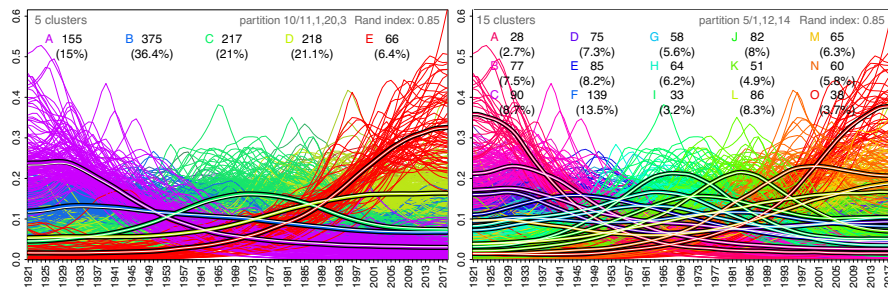


Fig. 1 Best partitions corresponding to the cluster number candidates $k = 5, 15$. The best partition maximizes the average Rand index of agreement with all the other partitions selected for k .

singles out more defined macro-dynamics that are in general nested in the basic patterns or, in a few cases, intermediate between them (Fig. 3).

To compare KBS to LDA findings, topic groups and word samples from hot/cold topics are illustrated in Figures 4 and 5. Topic trend analysis is able to identify significant trends just for a minor part of the overall topics, the rest showing an amorphous and non-interpretable trend (Fig. 4). Such inhomogeneity is even more remarkable considering that group size of topics is a little more than halved compared to that of curve clusters (30 vs. 15, even if word membership to topics is not unique). Still, a relevant number of words that most or strongly characterize the hot/cold topics (with the highest probability of membership) show a trend that goes in the opposite direction to that expected for the topic they represent (Fig. 5).

Interpretation of the history of ideas from KBS is out of the scope of this paper (only macro-dynamics of the finest partition are shown in Fig. 3 without the word list). We just note that LDA should lead to a greater interpretability since it can benefit from information on the co-presence of words in texts. However, it is common practice to base the interpretation on the first few (e.g., 5) words of the topic without integrating the information content of most words in a complete and consistent way. (Indeed, distribution of membership probability is generally highly right skewed.) Whereas, in KBS, interpretation involves all the group words, as they can represent a general theme but differentiated into sub-topics, different approaches and research fields. (Words are ordered according to an agreement index whose relatively low values help in identifying any word belonging to the group less firmly.)

5 Conclusions

This paper focus on the comparison between topic modelling and curve clustering with regard to temporal pattern detection. Topic trend analysis is able to identify significant trends just for a minor part of the overall topics. Even for the found hot/cold topics, the trend of a relevant number of representative words goes in the opposite direction to that expected (increasing/decreasing). However, one argument

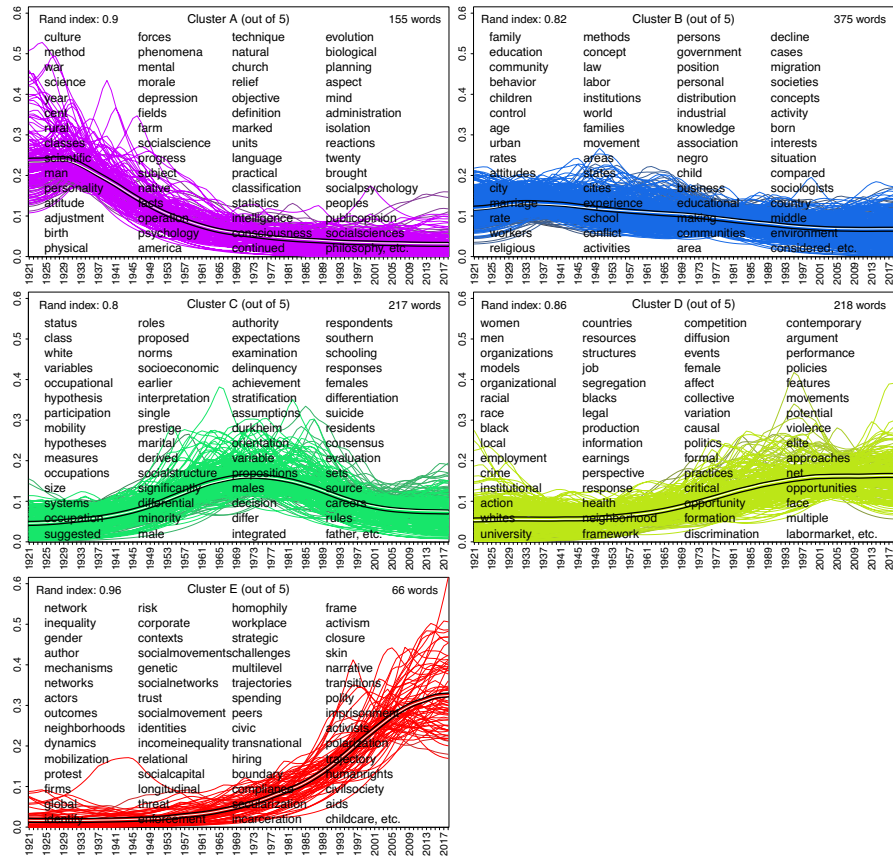


Fig. 2 Clusters of the 5-group partition. The subset of group words transcribed are ordered according to both popularity and individual (multiple) Rand index (both from highest to lowest).

in favor of LDA should be the greater interpretability due to the word co-presence in texts. Also on this regard, however, interpretation is commonly based on the first very few words of each topic without integrating the information content of most words. On the opposite side, clear trends are obtained by KBS at any level of word partitioning, from the roughest to the finest. Of course, some degrees of freedom are left to the analyst (but in the end decided together with the experts in the field): the level of partitioning (cluster number selection) and what aspects to privilege—popularity and/or synchrony—in combining word curves (type of normalization). Interpretation involves all the group words, as they can represent a general theme but differentiated into sub-topics, different approaches and research fields. In conclusion, we think that interpretation could be improved through a transversal reading—along the macro-dynamics discovered by KBS—of LDA topics in order to reintegrate the information of all or almost all the topic words and, thus, discover possible sub-topics or different historical moments of topic evolution.

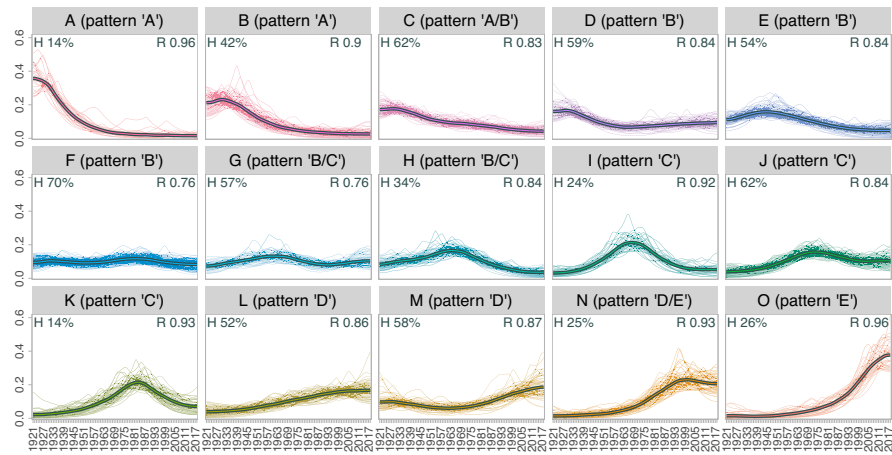


Fig. 3 Clusters of the 15-group partition ordered according to the five patterns of the 5-group partition ('A', 'B', 'C', 'D', 'E') and some intermediates. High-frequency words (H) % and (multiple) Rand index (R) are also indicated per group. Transparency of word curves is proportional to R.

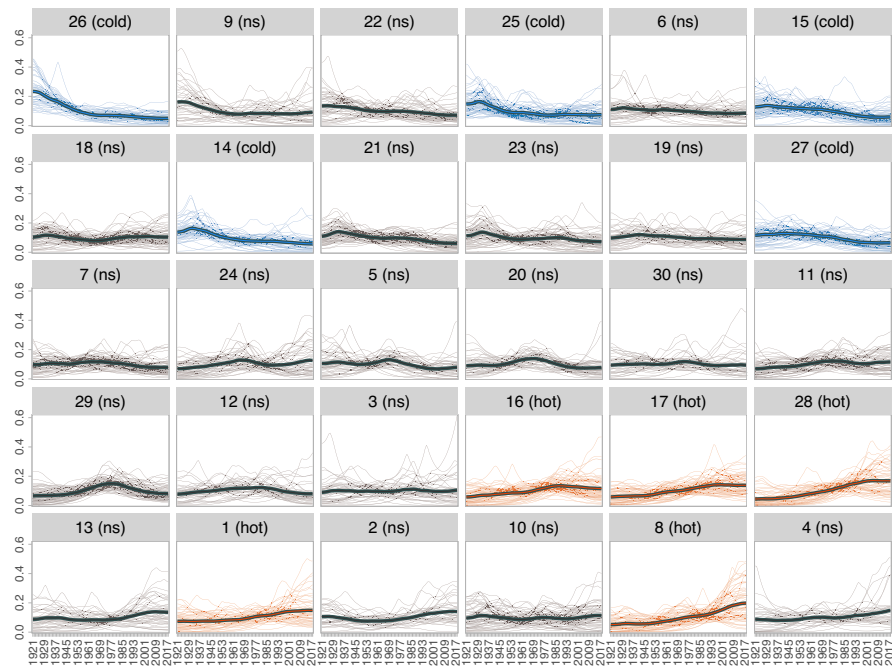


Fig. 4 The 30 groups of topic model set in chronological order. Hot/cold topics are highlighted. Transparency of word curves is proportional to topic membership probability.

References

1. Trevisani, M., Tuzzi, A.: A portrait of JASA: the History of Statistics through analysis of keyword counts in an early scientific journal. *Qual. & Quant.* **49**, 1287–1304 (2015)

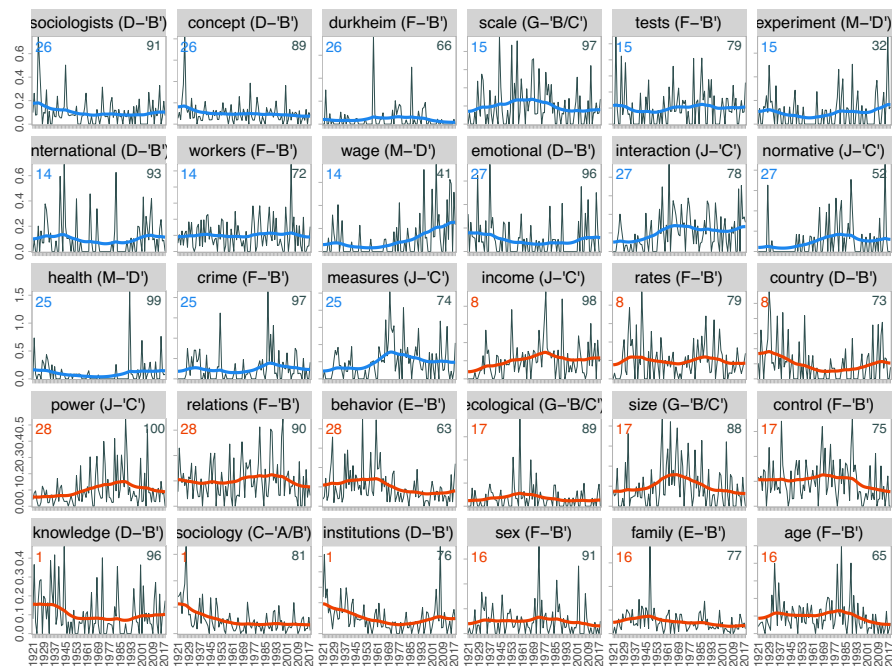


Fig. 5 Triplets of keywords that characterize the cold/hot topics and that do not follow the expected decreasing/increasing temporal trend. Panel label indicates 15-group partition cluster and 5-group partition pattern, topic number (top-left) and probability percentile (top-right) are also shown.

2. Trevisani, M., Tuzzi, A.: Learning the evolution of disciplines from scientific literature. A functional clustering approach to normalized keyword count trajectories. *Knowl.-Based Syst.* **146**(3), 129–141 (2018)
3. Trevisani, M., Tuzzi, A.: Chronological corpora curve clustering: From scientific corpora construction to knowledge dynamics discovery through word life-cycles clustering. *MethodsX* **5**, 1576–1587 (2018) pp. 105–129. Springer, Cham (2018)
4. Trevisani, M.: Functional Data Analysis and Knowledge-Based Systems. In: Tuzzi, A. (eds.) *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences*, pp. 167–187. Springer, Cham (2018)
5. Giordan, G., Saint-Blancat, C., Sbalchiero, S.: Exploring the History of American Sociology through Topic Modelling. In: Tuzzi, A. (eds.) *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences*, pp. 45–64. Springer, Cham (2018)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. *J. of Mach. Learn. Res.* **3**, 993–1022 (2003)
7. Blei, D.M., Lafferty, J.D.: Topic models. In: Sahami, A., Srivastava, M. (eds.) *Text Mining: Theory and Applications*, pp. 71–93. Taylor and Francis (2009)
8. Griffiths, T., Steyvers, M.: Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* **101**, 5228–5235 (2004)
9. Grn, B., Hornik, K.: Topicmodels: An R Package for Fitting Topic Model. *J. of Stat. Softw.* **13**, 1–30 (2011)
10. Sbalchiero, S., Tuzzi, A.: Whats Old and New? Discovering Topics in the American Journal of Sociology. In: Iezzi, D.F., Celdardo, L., Misuraca M. (eds.) *Proceedings of 14th International Conference on statistical analysis of textual data*, pp. 724–732. UniversItalia Editore (2018)

Classifying the Willingness to Act in Social Media Data: Supervised Machine Learning for U.N. 2030 Agenda

Classificare la volontà di agire nei dati dei Social Media: Supervised Machine Learning per l'Agenda 2030 delle Nazioni Unite

Andrea Sciandra, Alessio Surian and Livio Finos

Abstract In 2015, the United Nation General Assembly adopted the 2030 Agenda for Sustainable Development and its 17 Sustainable Development Goals aiming at ending all forms of poverty, fighting inequalities and tackling climate change. We collected data about the 2030 Agenda from May 9th to November 9th, 2018. The aim of this work is to obtain a classification of each tweet in the corpus according to the “Information” - “Action” categories, in order to detect whether a tweet refers to an event or it has only an informative-disclosure purpose. Explicit intention to act or inform had been captured by hand coding of a randomly selected sample of tweets and then the classification had been extended to the whole corpus through a supervised machine learning method.

Abstract L'assemblea generale delle Nazioni Unite adottò nel 2015 l'Agenda 2030 per lo sviluppo sostenibile e i suoi 17 obiettivi (SDG) per eliminare ogni forma di povertà, combattere le disuguaglianze e affrontare il cambiamento climatico. Abbiamo raccolto dati da social media sull'Agenda 2030 dal 9/5 al 9/11/2018 con lo scopo di ottenere una classificazione di ogni tweet secondo le categorie di “informazione” e “azione”, per capire se si riferissero a eventi specifici o avessero solo scopo informativo. Attraverso la codifica manuale di un campione casuale di tweet e all'uso di un metodo di apprendimento automatico supervisionato abbiamo esteso la classificazione a tutto il corpus di testi.

Key words: machine learning, textual data analysis, sustainable development goals

¹ Andrea Sciandra, STAR.Lab - Socio Territorial Analysis and Research, University of Padova; email: andrea.sciandra@unipd.it

Alessio Surian, Department of Philosophy, Sociology, Education and Applied Psychology (FISPPA), University of Padova; email: alessio.surian@unipd.it

Livio Finos, Department of Developmental Psychology and Socialisation, University of Padova; email: livio.finos@unipd.it

Introduction

In 2015, the United Nation General Assembly adopted the 2030 Agenda for Sustainable Development and its 17 Sustainable Development Goals (SDGs)² aiming at ending all forms of poverty, fighting inequalities and tackling climate change. The Agenda is a plan of action including 169 targets, National agencies and monitoring actions. At the international level, core actors include the United Nations Development Programme, UNDP, and the United Nations Environment Programme, UNEP.

In the monitoring the 2030 Agenda for Sustainable Development it seems particularly relevant Ulrich Beck's conceptual framework for the study of local responses to global risks, and particularly climate change. Decentralized action often studied in relation to national contexts. According to Beck et al. (2013) this approach falls short in relation to the global nature of this risk. Studies should be able to relate climate risk to the generative process encouraging the forming of cosmopolitan communities, i.e. facing a global risk should be mapped also in terms of cultural and political changes, and especially of those changes advocating a global perspective. Therefore, Beck et al. (2013) provide a theory concerning the formation of an operational capacity to face the climate change. Beck suggests the category of "emancipatory catastrophism" as a way to negotiate a third option in between the dichotomy that opposes an utopian approach, i.e. investing in "green" technology vs an apocalyptic approach, i.e. claiming a state of emergency and focusing on local conflicts. Emancipatory catastrophism implies a recognition of the limits of current knowledge and investing in a medium term scenario (the next two generations), halfway between the short term horizon of political urgency and the long term span of technical approaches.

As the 2030 Agenda is a plan of action, this paper intends to outline an interpretative schema by mapping the categories of information and action emerging in social media. It seems particularly interesting to understand how and to what extent people and organizations are playing a more active role in shaping the process of responding locally and internationally to climate change.

Data collection

Data about the 2030 Agenda had been collected through Twitter API using `twitterR` R package with OAuth authentication from May 9th to November 9th, 2018. By an access token associated to a Twitter app, we issued a search based on a text string

² The 17 SDGs: 1. No Poverty, 2. Zero Hunger, 3. Good Health and Well-being, 4. Quality Education, 5. Gender Equality, 6. Clean Water and Sanitation, 7. Affordable and Clean Energy, 8. Decent Work and Economic Growth, 9. Industry, Innovation and Infrastructure, 10. Reduced Inequality, 11. Sustainable Cities and Communities, 12. Responsible Consumption and Production, 13. Climate Action, 14. Life Below Water, 15. Life on Land, 16. Peace and Justice Strong Institutions, 17. Partnerships to achieve the Goals.

("agenda2030"). The resulting dataset for this study consists in a corpus of $N = 209216$ tweets. The Twitter streaming API allows users to gather up to 1% of all tweets that pass through the service at any time. Twitter users of course do not represent people as they are a particular sub-set and we can't assume that accounts and users are equivalent (Boyd & Crawford, 2012). We didn't look at the representativeness issue, because we were interested in empirically testing these methods, instead of providing a representative distribution of this phenomenon. However, some authors have shown that these limits can be overcome by improving the sample population coverage (Sampson et al., 2015), e.g. by splitting the same keywords across multiple crawlers and, in wider terms, a social media analysis may capture population attitudes and behaviours, even if the characteristics of the users do not reflect the characteristics of the full population (Ceron et al., 2017).

The increasing availability of digitized text, especially from Social Media, offers enormous opportunities for social scientists, even though this kind of data contain lot of noise. In fact, text fragments coming from social media can be considered as off-topic for the purpose of the analysis very frequently. It must be stressed that commonly off-topic texts don't use completely different words from the text of the training set, so classifiers will attribute as outcome this category with high probability and very rarely the true semantic category (Ceron et al., 2017).

Pre-processing

First of all, we carried out a language detection, as we decided to focus on the Spanish language tweets (about 132000), for specific research interests and to achieve a certain degree of heterogeneity, including a large European country and most of central and south America. Anyway, since we intend to use Supervised Machine Learning (SML) techniques, we must remember that these methods are completely independent of the language of the text (moreover, SML improves estimates and there is no a priori set of categories).

The pre-processing phase involves the following steps:

- Identification and homogenization of retweets by comparison between texts with Levenshtein distance (RThound function - TextWillaer R package); after these procedure, the retweets count is equal to 84.7%, a very high share.
- Text cleaning, in order to normalize text encoding, html, emoticons, punctuation, etc. Therefore, all the URLs appearing in the corpus had been coded in a single word (wwwurlwww) replacing all the links. If an emoticon had been recognized, it'd been coded with a word that specify the kind of emoticon (eg ";)") had been recoded into "emotewink").
- Stopwords removing and word stemming in Spanish language. The initial corpus had 39193 type and 2726049 token (type/token ratio: 14%), while after pre-processing, removing stopwords and stemming we had 1482492 token and 35989 type (type/token ratio: 2%).

Figure 1 shows a wordcloud of the most frequent words.

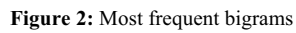
Figure 1: Most frequent words (wordcloud)

A textual analysis allowed us to include some variables related to SDGs hashtag and also the tf-idf weighting (Salton & Buckley, 1988), to show how important a word is to distinguish “Information” and “Action” tweets in our corpus. In particular, some simple text mining procedures allowed us to create the document-term matrix we used for the classifications. We worked within the "bag of words" framework, as only tf-idf weighting was applied to the words³ and we don't assume any Natural Language Processing rules nor do we want to use ontological dictionaries. We also analyzed the main multiword expressions by their frequencies as we tokenized adjacent words into n-grams. The aim of this procedure was to identify some patterns in our corpus, potentially useful for the hand-coding phase, by visualizing some recurrent relationships among words simultaneously (Figure 2). Moreover, in very few cases we recoded some n-grams because they were equivalent to acronyms (e.g.: "objetivos desarrollo sostenible" has been coded as "ods", a very frequent acronym and hashtag).

Since we had a big sparse document-term matrix (35989 terms), we decided to allow maximal sparsity at 99.9%, or 0.01% in relation to document frequency. The resulting matrix contains only 1745 terms, since we removed all the terms which have at least a 99.9% of empty elements (terms occurring 0 times in a document).

³ Term frequency was normalized by the number of occurrences of all terms of each tweet.

Classifying the Willingness to Act in Social Media Data



Training set

The aim of this work is to obtain a classification of each tweet in the corpus according to the "Information" - "Action" categories, in order to detect whether a tweet refers to an event or it has only an informative-disclosure purpose. Explicit intention to act or inform had been captured by hand coding of a randomly selected sample of 1584 tweets and then the classification had been extended to the whole corpus through a supervised machine learning method. The data had been hand coded by the three authors.

1.1 Hand coding

After randomly selecting a subset of 2000 (1584 unique), 2030 Agenda tweets were tagged manually in 2 (4) dimensions:

- "Information" (split into "General Information" or "Information with data");
- "Action" (split into "Action: Policies/Political Actors" or "Territorial Action").

The Information categories have been hand coded according to the following recommendations:

- General Information, i.e. tweets for merely informative purposes, which also include an invitation to follow a political event; this type is called "general information";
- "Information with data" to highlight the informative tweets where there are some data or the data of a report are mentioned.

Otherwise, the dimension of the action should refer to creating something or at least a strong encouragement to do so. In particular, this category had been split into:

- "Action: Policies/Political Actors", i.e. tweets talking about national and local government as well as NGO actions or specific public policies;

- "Territorial Action": concrete projects and initiatives are cited, in other words, real actions on the territory. For "Territorial Action" a good indicator could be the word *taller* (laboratory/workshop), as well as *acción*.

The 4 dimensions are mutually exclusive and are distributed as follows:

- General Information: 72.35%
- Information with data: 10.48%
- Action: Policies/Political Actors: 9.66%
- Territorial Action: 7.51%

The agreement rate between coders was 90.06% and disagreement cases had been corrected following a discussion between the three coders.

Supervised Machine Learning

Among the many, we focused on Gradient Boosting method (Friedman, 2002, R implementation by Chen et al., 2019) for its flexibility and good performances. The Gradient Boosting method results to be particularly suitable for models characterized by sparse features since it produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

We chose to use all the terms of the reduced document-term matrix as features, weighted by their tf-idf. Some more variables collected through Twitter API were added as features (retweet count, favourite count, etc.).

The model is set to predict the four (or two) dimensions – i.e. classes – and is fitted to maximize the accuracy of the prediction.

For the estimate of the proportion of each dimension – i.e. topic/class – we use the estimated probability of each tweet, so that each tweet contribute to the estimate of each dimension.

Results

A summary of the performances of the model with four and two dimensions is reported in Table 1. Despite not optimal, the indices demonstrate the ability to predict the dimension (i.e. the topic) of the tweet.

Table 1 Performance indices of prediction models

<i>Four Dimensions model</i>			<i>Two Dimensions model</i>		
Accuracy	0.810		Accuracy	0.864	
	Precision	Recall		Precision	Recall
Info-Generic:	0.805	0.974		0.869	0.988

	<i>Four Dimensions model</i>	<i>Two Dimensions model</i>
Info-Data:	0.805	0.545
Action-politics:	0.750	0.300
Territ.-Action:	1.000	0.217

The importance of the most relevant features and words in the model (4-dimensions) are shown in Figure 3.

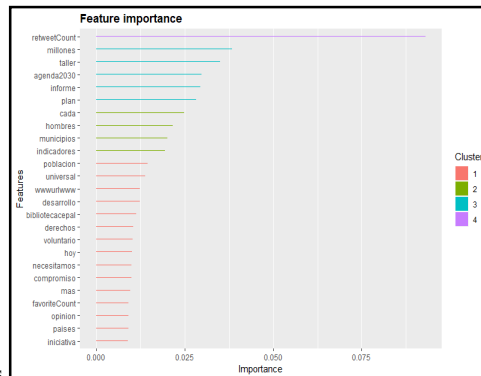


Figure 3: Importance of top 25 features

The estimate of dimension's proportion over the whole dataset (131029 documents) is given in Table 2.

Table 2 Estimated percentage of the four dimensions (131029 documents)

<i>Dimension</i>	<i>Percentage</i>
General Information	75.74%
Information with data	8.74%
Action: Policies/Political Actors	9.16%
Territorial Action	6.36%

Conclusions

Beck et al. (2013) identified cosmopolitan forms of action in (1) environmental associations (including green consumerism); (2) sustainable cities networks, focusing on sustainable territorial planning; and (3) innovation consortia (including “green” entrepreneurs and/or products). These “cosmopolitan communities of climate risk” seem quite underrepresented in our sample. Actions as such are addressed only in marginal ways. Through hand coding we found very few tweets that talked about concrete actions, potentially falling into the clicktivism/slacktivism (Morozov 2009) phenomenon, i.e. many users retweeted but actually there is little concrete commitment (e.g. laboratory/workshop as territorial actions) or, in general, a willingness to act.

Overall, our Classification Supervised Model has produced satisfactory results, both for 2 and 4 categories, showing the importance of the variables deriving from textual analysis and those coming directly from Twitter.

Although we are aware of the limitations of this kind of studies (representativeness, noise, bots, etc.), we believe that probabilistic short-term models applied to the empirical observation of human behaviour in large datasets mined from social media could extract social knowledge, representing a great opportunity for social scientists (Lauro et al., 2017).

References

1. Beck, U., Blok, A., Tyfield, D., Zhang, J.Y.: Cosmopolitan Communities of Climate Risk: Conceptual and Empirical Suggestions for a New Research Agenda, *Glob. Netw.*, 13(1), 1-21 (2013)
2. Boyd, D. & Krawford, K. (2012). Critical questions for big data. *Inf., Comm. and Society*, XV(5), 662–679. doi:10.1080/1369118X.2012.678878.
3. Ceron, A., Curini, L., Iacus, S.M.: *Politics and Big Data: Nowcasting and Forecasting Elections with Social Media*. Routledge, New York (2017)
4. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y.: *xgboost: Extreme Gradient Boosting*, R package version 0.81.0.1, 1-4 (2019)
5. Friedman, J.H.: Stochastic gradient boosting. *Comput. Stat. & Data Anal. (Nonlinear methods and data mining archive)*. 38(4), 367-378 (2002)
6. Lauro, N.C., Amaturio, E., Grassia, M.G., Aragona, B., Marino, M. (Eds.): *Data Science and Social Research: Epistemology, Methods, Technology and Applications*. Springer, Heidelberg (2017)
7. Morozov, E.: *The Brave New World of Slacktivism*. *Foreign Policy*. Available online: <https://foreignpolicy.com/2009/05/19/the-brave-new-world-of-slacktivism> - accessed 15/03/2019 - (2009)
8. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Inf. Process. & Manag.* 24(5), 513-523 (1988)
9. United Nations General Assembly: *Transforming our world: the 2030 Agenda for Sustainable Development*, Resolution 70/1 (2015).

Classification of spatio-temporal point pattern in the presence of clutter using K -th nearest neighbour distances

Classificazione dei processi puntuali spazio-temporali basata sulla distanza dal K -mo vicino più vicino

Siino Marianna, Francisco J. Rodríguez-Cortés, Jorge Mateu, Giada Adelfio

Abstract In a point process spatio-temporal framework, we consider the problem of features detection in the presence of clutters. We extend the methodology of Byers and Raftery (1998) to the spatio-temporal context by considering the properties of the K -th nearest-neighbour distances. We make use of the spatio-temporal distance based on the Euclidean norm where the temporal term is properly weighted. We show the form of the probability distributions of such K -th nearest-neighbour distance. A mixture distribution, whose parameters are estimated with an EM algorithm, is used to classify points into clutters or features. We assess the performance of the proposed approach with a simulation study, together with an application to earthquakes.

Abstract *Nell'ambito dei processi puntuali spazio-temporali, abbiamo affrontato il problema relativo all'identificazione di aggregazione di punti in presenza di di eventi di fondo. Abbiamo esteso la metodologia proposta da Byers and Raftery (1998) nel contesto spazio-temporale considerando le proprietà (quali la distribuzione di probabilità) della distanza dal K -mo vicino più vicino. In particolare, abbiamo fatto uso della distanza Euclidea (dove la parte temporale è opportunamente pesata con un fattore di scala) mostrando la distribuzione di probabilità di questa distanza. Un approccio basato sulla misture di distribuzioni è stato utilizzato per classificare i punti nelle due rispettive categorie, utilizzando il metodo di risoluzione EM. Con-*

Siino Marianna

Istituto Nazionale di Geofisica e Vulcanologia, Centro Nazionale Terremoti, Rome, Italy, e-mail: marianna.siino@ingv.it

Francisco J. Rodríguez-Cortés

Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia e-mail: frrodriguez@unal.edu.co

Jorge Mateu

Departament of Mathematics, University Jaume I, Castellón, Spain e-mail: mateu@uji.es

Giada Adelfio

Dipartimento di Scienze Economiche, Aziendali e Statistiche,
Università degli Studi di Palermo, Palermo, Italy e-mail: giada.adelfio@unipa.it

siderando diversi scenari, abbiamo verificato il comportamento del metodo proposto. Inoltre, abbiamo considerato un'applicazione in ambito sismico per la classificazione dei terremoti in eventi di fondo e indotti.

Key words: Clutter; Earthquakes; EM algorithm; Features; Mixtures; Nearest-neighbour distances; Spatio-temporal point patterns.

1 Introduction

One of the most important research fields of spatio-temporal data-mining is the identification of features (clusters) of events. In particular, features are defined as subgroups of events in constrained spatio-temporal volumes with a higher density than other events outside the spatio-temporal windows (called background, noise or clutter events). The identification of such spatio-temporal features may yield insight for many applications. In practice, many geographical phenomena (e.g. earthquakes, disease cases, crime data, forest fires) are modelled as spatio-temporal events, and the detection of features is used to study the evolution of the phenomena, to reveal space or time anomalies and spatio-temporal hotspots. However, the detection of features is a challenge problem for the complexity caused by the time-space coupling and the noise interference.

For point processes, the corresponding problem has been widely addressed from a spatial point of view (Allard and Fraley, 1997; Byers and Raftery, 1998; Dasgupta and Raftery, 1998; Illian et al, 2008). For instance, Byers and Raftery (1998) estimated and removed the clutter without making any assumptions about the shape or number of features. Their method of detection is based on the distance to the K -th nearest-neighbour of a point in a spatial process. The observed distances are modelled as a mixture distribution of distances coming from clutter and feature points, the parameters of which are estimated by an EM algorithm.

In this paper, we use spatio-temporal distances obtained as a weighted version (with respect to the temporal component) of the Euclidean distance (Demattei and Cucala, 2010). Moreover, it is shown that the distribution of the K -th nearest-neighbour based on the previous distance follows an inverse Gamma distribution under the homogeneous Poisson assumption. The spatio-temporal K -th nearest-neighbour distance is analysed through a mixture model formulation of the corresponding distance densities coming from the clutter and feature events. The corresponding parameters associated to the two density distributions in the mixture model formulation are estimated using an expectation-maximisation (EM) algorithm.

A simulation study is carried out to assess the performance of the proposed classification method. Our method is also compared with the results obtained with the spatial methodology of Byers and Raftery (1998) in terms of sensitivity, specificity and accuracy. Finally, we present a seismic application, identifying noise and feature events in the seismic sequences occurred in California (near the Landers town, in 1992).

2 Methodology

We consider a spatio-temporal point process with no multiple points as a random countable subset X of $\mathbb{R}^{d-1} \times \mathbb{R}$, where a point $(\mathbf{u}, s) \in X$ corresponds to an event at $\mathbf{u} \in \mathbb{R}^{d-1}$ occurring at time $s \in \mathbb{R}$. We observe n events $\{(\mathbf{u}_i, s_i)\}_{i=1}^n$ of distinct points of X within a bounded spatio-temporal region $W \times T \subset \mathbb{R}^{d-1} \times \mathbb{R}$, with volume $|W| > 0$, and with length $|T| > 0$ where $n \geq 0$ is not fixed in advance. We assume that the point process X is stationary and isotropic. Let $(\mathbf{u}, s) = (u_1, u_2, \dots, u_{d-1}, s)$, and $(\mathbf{v}, l) = (v_1, v_2, \dots, v_{d-1}, l)$ be points of a spatio-temporal homogeneous Poisson process in $W \times T \subset \mathbb{R}^{d-1} \times \mathbb{R}$. Following Demattei and Cucala (2010) and given $d = 2$, we consider the spatio-temporal Euclidean distance given by

$$D^{ST_E}((\mathbf{u}, s), (\mathbf{v}, l)) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \rho^2 |s - l|^2}, \quad (1)$$

that is a mixture of a spatial and temporal distances where ρ is a scaling coefficient for the temporal part. The close form of the probability distribution function for the K -th nearest-neighbour Euclidean distance (Siino et al, 2019), under a homogeneous spatio-temporal Poisson process is given by

$$D_K^{ST_E} \sim \Gamma^{1/d} \left(K, \frac{\pi^{d/2} \lambda}{\Gamma(d/2 + 1) \rho} \right). \quad (2)$$

The maximum likelihood estimate (MLE) of the rate of the process is given by $\hat{\lambda} = K / (\alpha_d \sum_{i=1}^N \gamma_i^d)$ where the γ_i are the observations of $D_K^{ST_E}$, N is the sample size and $\alpha_d = (\pi^{d/2}) / (\Gamma(d/2 + 1) \rho)$.

As in Byers and Raftery (1998) and Mateu et al (2007), we assume to have two types of processes to be classified, and model the K -th nearest-neighbour distances through a mixture of the corresponding K -th nearest-neighbour distances coming from the clutter and feature, which are two superimposed spatio-temporal Poisson processes. We then suppose that the distribution of the $D_K^{ST_E}$ is roughly a mixture of distributions

$$D_K^{ST_E} \sim q \Gamma^{1/d}(K, \alpha_d \lambda_1) + (1 - q) \Gamma^{1/d}(K, \alpha_d \lambda_2), \quad (3)$$

where λ_1 and λ_2 are the intensities of the two homogeneous spatio-temporal Poisson point processes (clutter and feature) and q is the constant which characterises the postulated distribution of the $D_K^{ST_E}$. The corresponding parameters associated to this mixture are estimated using an expectation-maximisation (EM) algorithm, where in the expectation step we use the close form provided by an inverse Gamma distribution.

3 Simulation study

A simulation study is carried out to assess the performance of the proposed methodology in terms of detection of features in a spatio-temporal setting. The spatio-temporal classification procedure proposed in this paper is compared with the method based on the spatial K -th nearest-neighbour distance in Byers and Raftery (1998), named $M_{spatial}$, as if we ignore time. The spatio-temporal window is set as

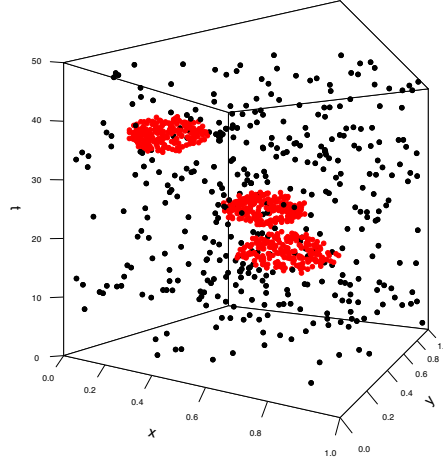


Fig. 1: Simulated scenarios with $n_c = 400$ clutter points from a homogeneous Poisson point process in a spatio-temporal window $[0, 1]^2 \times [0, 50]$. There are three ellipsoids ($n_{clusters} = 3$) with semi-axis $a = 0.2$, $b = 0.15$ and $c = 3.5$ with volume 0.43, each with $n_{fc} = 200$ points, and $n_f = 600$.

$W \times T = [0, 1]^2 \times [0, 50]$, where the time range is greater than the spatial one to have a different scale than the spatial window. In the spatio-temporal window, the clutters and features are simulated from two different processes. The clutter points are taken from a spatio-temporal homogeneous Poisson process with size $n_c = \{400, 1000\}$. There are three clusters ($n_{cluster} = 3$) and the number of feature points for each cluster is $n_{fc} = \{200, 400\}$, so the total number of feature points is $n_f = n_{fc} \times n_{cluster}$. Figure 1 shows a simulated point pattern for $n_c = 400$ clutter points and a number of feature points per cluster of $n_{fc} = 200$, with three ellipsoids.

We report results for $K = \{5, 10\}$ and $\rho = \{0.02, 0.5, 1\}$. Parameter ρ rescales the temporal distance, such that the weight of the temporal term changes accordingly. When $\rho = 1$, the methodology is equal to $M_{spatial}$ in three dimensions (Byers and Raftery, 1998). Instead, when $\rho = 0.02$, it corresponds to simulate the points in the unit cube. For each scenario, 100 point patterns are simulated as described above. The evaluation and comparison of the classification procedures with respect to the

different settings are done in terms of the true positive rate (TPR), the false positive rate (FPR) and the accuracy (Stehman, 1997). The results are shown in Table 1.

Our method always outperforms $M_{spatial}$, in terms of TPR, FPR and accuracy. Decreasing the value of ρ , so making closer the unit measurements of the space and time dimensions, the TPR, the specificity (1-FPR) and accuracy increase. In general, we observe that changing the value of K (from 5 to 10) the overall performance is comparable. Moreover, as expected, increasing the number of clutter points in $W \times T$ ($n_c = 1000$), the results are in general slightly worse than the cases with $n_c = 400$. For further simulation results and comparisons see Siino et al (2019).

Table 1: Results in terms of true positive rate (first row), false positive rate (second row) and accuracy (third row) in percentage over 100 simulated point patterns when the spatio-temporal feature points are three ellipsoids. The total number of clutter points is $n_c = \{400, 1000\}$. $n_{fc} = \{200, 400\}$ indicates the number of points for each feature, K is the K -th nearest-neighbour distance ($K = \{5, 10\}$) and $\rho = \{1, 0.5, 0.02\}$ is the weight for the temporal term in the spatio-temporal distance. D^{STE} refers to the Euclidean space-time distance. $M_{spatial}$ indicates the results obtained with the spatial method of Byers and Raftery (1998), neglecting time.

		$n_c = 400$				$n_c = 1000$			
		$K = 5$		$K = 10$		$K = 5$		$K = 10$	
n_f	ρ	D^{STE}	$M_{spatial}$	D^{STE}	$M_{spatial}$	D^{STE}	$M_{spatial}$	D^{STE}	$M_{spatial}$
200	1	97.72	96.24	96.16	97.80	97.51	92.87	96.50	95.26
		8.82	26.89	11.89	26.42	9.20	33.07	10.43	27.44
		95.11	86.99	92.94	88.12	93.32	76.66	92.17	81.07
	0.5	98.97	-	98.03	-	98.59	-	98.20	-
		6.64	-	8.33	-	6.76	-	7.59	-
		96.73	-	95.49	-	95.25	-	94.58	-
	0.02	99.83	-	99.96	-	99.54	-	99.88	-
		4.80	-	5.72	-	4.49	-	5.05	-
		97.98	-	97.69	-	97.02	-	96.80	-
	400	99.08	97.75	98.31	98.78	98.79	95.64	98.09	97.04
		7.50	26.00	9.72	26.43	6.65	26.93	8.12	25.01
		97.43	91.81	96.30	92.48	96.32	85.38	95.27	87.02
400	0.5	99.62	-	99.26	-	99.43	-	99.05	-
		5.81	-	7.26	-	5.37	-	6.39	-
		98.26	-	97.63	-	97.25	-	96.58	-
	0.02	99.96	-	100.00	-	99.83	-	99.97	-
		4.46	-	5.28	-	4.04	-	4.61	-
		98.86	-	98.68	-	98.07	-	97.89	-

4 Application on California earthquakes

In this section, the proposed method is applied to seismic data. Since an earthquake can be viewed as a spatio-temporal pattern, the identification of clustered earthquakes provides key information on seismic dynamics. Well-studied statistical models are based on the idea that the seismicity can be considered as the sum of “background” earthquakes (caused by tectonic loading) and “triggered” earthquakes (Ogata, 1988; Adelfio and Chiodi, 2015). To describe the seismicity of an area in space, time and magnitude domains, sometimes it is useful to study separately the features of *independent* events and *triggered* ones. At this regard, the proposed method based on the EM algorithm allows the identification of these two main components, such that the background seismicity is related to the long-term analysis, and the triggered one for sequence identification.

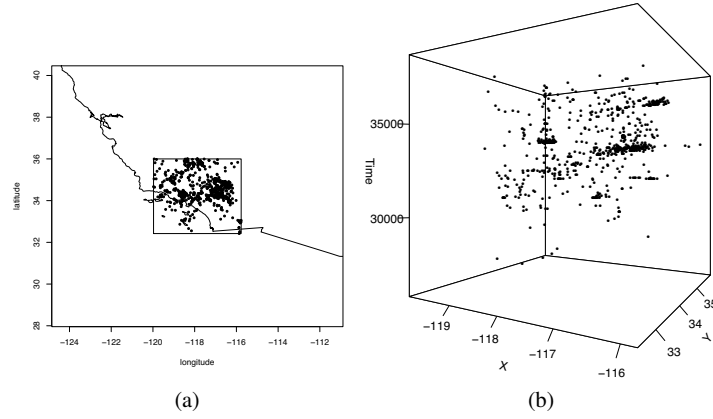


Fig. 2: 2D and 3D-plots of the earthquake near Landers city, California.

Our goal is to detect spatio-temporal features (clusters) of earthquakes that occurred in California choosing a spatial window around high seismic areas. The subset of the earthquake catalogue in California refers to a study area located between -120–115 N and 32–36 E, near the Landers town, where in June 25th, 1992 an event with magnitude 7.3 occurred, causing severe damage to the area directly surrounding the epicenter. A total number of 804 were observed with magnitude at least 3.5 from 1968 to 2012 (see Figure 2).

The value of ρ is set as the ratio between the maximum spatial distance over the maximum temporal distance observed between the events, so $\rho_{california} = 0.042$. The entropy measure ($S = \sum_{i=1}^n \delta_i \log(\delta_i)$ where δ_i are the probabilities of being in the feature group) for each value of K is reported, Figure 3a. We selected the value of $K = 25$ since after this value the entropy measure can be considered constant. The histograms of the selected K -th nearest-neighbour distance based on the Euclidean

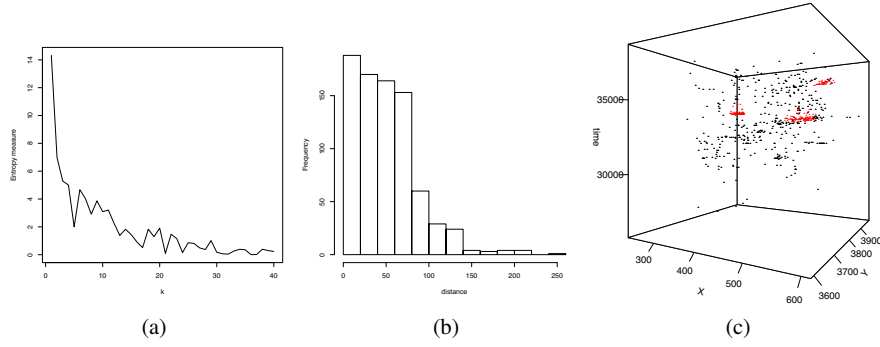


Fig. 3: (3a) Entropy measure changing the K -th value of the K -th nearest-neighbour using D^{STE} . (3b) Histogram of the D^{STE} distance to the 25-th nearest-neighbour. (3c) Detected feature and clutter points with $\rho_{california}$ in D^{STE} .

weighted distance is in Figure 3b. The results applying the EM classification procedure are reported in Figure 3c, where we can see that the feature points are clearly identified. The total number of clutter and feature points were $n_c = 477$ and $n_f = 327$ and three spatio-temporal features are identified.

5 Conclusions

In this paper, we present a classification method for identifying regions with higher point densities (features) in a spatio-temporal context extending the procedure of Byers and Raftery (1998) that is based on the spatial K -th nearest-neighbour distance. We use the weighted Euclidean spatio-temporal distance where the temporal term is scaled to account for the space-time coupling. Based on these results, a mixture model formulation for the K -th nearest-neighbour distance of clutter and feature points is considered to perform a binary classification using an iterative EM algorithm. With a simulation study with clusters in space and time, the proposed methodology is compared with the spatial version of Byers and Raftery (1998). The comparison is done in terms of the true positive rate, the false positive rate and the accuracy. In general, when weighting the temporal component in the distance measure, the results of the classification improve. In comparison to the spatial method, our methodology outperforms the other one. The analysis of the California sequences, to identify background and triggered events, shows its utility and wide application.

References

- Adelfio G, Chiodi M (2015) Flp estimation of semi-parametric models for space–time point processes and diagnostic tools. *Spatial Statistics* 14:119 – 132
- Allard D, Fraley C (1997) Nonparametric maximum likelihood estimation of features in spatial point processes using voronoi tessellation. *Journal of the American Statistical Association* 92(440):1485–1493
- Byers S, Raftery AE (1998) Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association* 93(442):577–584
- Dasgupta A, Raftery AE (1998) Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* 93(441):294–302
- Demattei C, Cucala L (2010) Multiple spatio-temporal cluster detection for case event data: An ordering-based approach. *Communications in Statistics - Theory and Methods* 40(2):358–372
- Illian J, Penttinen A, Stoyan H, Stoyan D (2008) *Statistical Analysis and Modelling of Spatial Point Patterns*, vol 70. John Wiley & Sons
- Mateu J, Lorenzo G, Porcu E (2007) Detecting features in spatial point processes with clutter via local indicators of spatial association. *Journal of Computational and Graphical Statistics* 16(4):968–990
- Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* 83(401):9–27
- Siino M, Rodríguez-Cortés FJ, Mateu J, Adelfio G (2019) Spatio-temporal classification in point patterns under the presence of clutter. Submitted
- Stehman SV (1997) Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment* 62(1):77–89

Modelling properties of high-dimensional molecular systems

La modellazione di sistemi molecolari ad alta dimensionalità

Debora Slanzi, Valentina Mameli and Irene Poli

Abstract The complexity of biological systems often causes scientists to generate large quantities of data, large in size and dimensionality, to build models and gain insights into the systems. In drug discovery, this problem is commonly encountered in the phase known as lead optimisation where the objective is to find a set of molecules with particular properties. With the aim of building and identifying these molecules, we propose a smart strategy to select a small set of molecular compositions that evolves in dimension and information in the experimental space. This strategy is based on a sequential design built with the predictive information achieved with statistical models for high dimensionality. The performance of the approach is evaluated in the study of the MMP-12 enzyme.

Abstract *La complessità dei sistemi biologici spesso porta alla generazione di grandi insiemi di dati, grandi per numerosità e dimensionalità, con l'obiettivo di costruire modelli per investigare aspetti della loro struttura. Nel settore di ricerca dedicato alla costruzione di nuove medicine, questo problema viene generalmente affrontato nella fase nota come lead optimisation, dove l'obiettivo è trovare un insieme di molecole che soddisfino alcune proprietà essenziali. Con lo scopo di costruire e identificare tali molecole, proponiamo una strategia intelligente per selezionare un insieme molto piccolo di composizioni molecolari che evolve in dimensione e informazione nello spazio sperimentale. Tale strategia si basa su un disegno sequenziale costruito con l'informazione predittiva derivante da modelli*

Debora Slanzi

Department of Management, Ca' Foscari University of Venice, San Giobbe, Cannaregio 873, Venice (IT) and European Centre for Living Technology, Dorsoduro 3911, Venice (IT), e-mail: debora.slanzi@unive.it

Valentina Mameli

Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, via Torino 155, Mestre (IT) and European Centre for Living Technology, Dorsoduro 3911, Venice (IT) e-mail: valentina.mameli@unive.it

Irene Poli

European Centre for Living Technology, Dorsoduro 3911, Venice (IT), e-mail: irenpoli@unive.it

statistici per alta dimensionalità. Questo approccio viene quindi valutato nello studio dell'enzima MMP-12.

Key words: high-dimensional systems, multi-objective optimisation, smart evolutionary strategies, predictive models.

1 Introduction

One of the main objectives of drug discovery research is to find a molecule that interacts with a pharmacological target (protein or pathway) to produce a therapeutic effect for a particular disease. Current drug design practices involve the screening of large chemical libraries, composed of thousands or millions of compounds, with the aim of identifying a candidate molecule with suitable properties, called the *lead molecule*. To model the relation between molecule properties and the set of candidate features affecting these properties, some aspects must be considered. The set of features is typically extremely large (of the order of several thousand) but experiments are costly and the number of molecules to be tested in laboratory is usually limited to a very small set. This kind of laboratory experimentation often produces “small high-dimensional data”, where the number p of variables (features) is much larger than the number n of observations (tests in laboratory). Thus, the sparsity of the information in such small high-dimensional data makes difficult, or inadequate, the standard methodological approaches. Furthermore, molecules must exhibit the required combination of efficacy and safety to potentially become a successful therapeutic. A set of pharmacokinetic properties, such as Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET properties) have to be predicted and optimised in order to minimise the risk of late-stage attrition and reduce the number of safety issues. This leads to multi-objective problems where multivariate modelling for high-dimensional settings should be taken into account to infer the information on the system. Considering all these issues, the modelling and the prediction for optimisation of the relations among the large set of features describing the molecules and their associated experimental properties can be a challenging problem. Methodologies for modelling and optimising systems in drug discovery include de novo designs [4], molecular docking [6] and quantitative structure-activity relationships [13]. These approaches have been successful in detecting the relevant information for discovering the molecules with optimal properties using computer based algorithms.

In this work we introduce a new procedure for drug discovery research by a smart approach to model, predict and optimise the properties of high-dimensional molecular systems for detecting those molecules that can be considered as lead molecules. We develop an evolutionary approach based on statistical models to extract information in small high-dimensional data. This approach, named *model-based evolutionary approach for optimisation*, combines the principles of the evolutionary paradigm with the inferential advantages of statistical models [1, 2, 10, 11]. Specif-

ically we generate and evolve a small set of data on which to estimate suitable models and use these models to make predictions on the next set of data to generate. Focus of the evolution is to detect the sample of data with which to estimate models for optimising the response of the system, i.e. the properties of the lead molecules. We test this approach to derive the best molecules for the analysis of the MatrixMetalloProteinase-12 (MMP-12), an enzyme that in humans is involved in several inflammation pathways [8].

The paper is organised as follows. In Section 2 we briefly present the model-based evolutionary approach for optimisation and we introduce the predictive models in which we will focus. Some key aspects of the multi-objective optimisation are also presented. Section 3 describes the characteristics of the data available for the MMP-12 study and presents some results. Finally in Section 4 we derive some concluding remarks.

2 The model-based evolutionary approach for optimisation

Consider a finite space \mathbf{X} , whose elements are the candidate sets of molecules (experimental points, the compositions to be tested in laboratory), $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)'$, with $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, $i = 1, \dots, N$, being the i -th vector of observations on the set of p variables representing the system. Define an objective function $f_k : \mathbf{X} \mapsto \mathbb{R}$ such as

$$Y_k = f_k(\mathbf{X}) + \varepsilon \quad (1)$$

where Y_k represents the response variable in the model and ε is a stochastic error term with a particular probability distribution. Optimisation consists of finding the best values of an objective function given a defined domain. In a maximisation task, the process of optimisation searches for the elements $\mathbf{x}^* \in \mathbf{X}$ such that $f_k(\mathbf{x}^*) \geq f_k(\mathbf{x})$ for all $\mathbf{x} \in \mathbf{X}$. Since we do not know the function f_k , we must estimate it using a training set of data from the elements of \mathbf{X} . We assume that the training set has a very small dimension with respect to the finite space \mathbf{X} , namely n , with $n \ll N$, and that it will be selected step by step during the process. A simple and general description of the strategy can be the following:

- A first set of elements $\mathbf{X}^1 = (\mathbf{x}_1, \dots, \mathbf{x}_{n_1})'$ with $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n_1$, is randomly chosen (or integrated with possibly prior information). Each element corresponds to a candidate molecule of the system under study. The experiment is conducted, and a n_1 -dimensional response vector $\mathbf{y}_k^1 = (y_{k,1}, \dots, y_{k,n_1})'$ is generated. Response vectors correspond to the molecular properties observed for the molecules selected and tested in laboratory in the current generation.
- Predictive models A_l , $l = 1, \dots, L$, are estimated on $[\mathbf{X}^1, \mathbf{y}_k^1]$ to model the objective function (response) on this first collected data.
- Predictions on the remaining subset of $(N - n_1)$ elements are derived, inferring the properties of all the molecules not tested in laboratory, being the test set on which to derive predictions. The molecules associated to the best predicted values

become the second population $\mathbf{X}^2 = (\mathbf{x}_1, \dots, \mathbf{x}_{n_2})'$, of experimental points to be tested.

Within this process, the information gathered from models is used to drive the evolution towards the optimal value. In fact, with this information we can identify the next generation of experimental points which evolves from one generation to the next by capturing the characteristics of the data testing a very small set of data. This iterative procedure of sampling, modelling and predicting, continues for M generations until a pre-defined number n of experiments is conducted, with $n_1 + n_2 + \dots + n_M = n < N$.

2.1 Models for prediction

If the objective function can be assumed to be linearly related to the predictors, then Eq. 2 is

$$Y_k = \mathbf{X}\beta + \varepsilon \quad (2)$$

where $\beta = (\beta_1, \dots, \beta_p)^T$ is the regression vector of unknown parameters associated to the p predictors. Usually molecular systems are characterised by small high-dimensional data where the number of predictors is much larger than the number of observations. Under this set-up, penalised regression procedures offer powerful methods to simultaneously estimate models and perform variable selection. Among these procedures, the most common is the least absolute shrinkage selection operator (Lasso); see [14].

Alternative to linear regression, the Neural Networks (NN) models can be considered for this structure of data. NN in fact are suitable models for dealing with data characterised by complex non-linear relationships and have become a popular tool for many applications in a wide range of fields including drug discovery research. The topology of a NN can be described as a collection of nodes, called neurons, which are arranged into ordered layers. A NN usually consists of input, hidden and output layers. With a single hidden layer, the dynamics of the information can be summarised by the following expression

$$\mathbf{Y} = f(\phi(\mathbf{X}, \mathbf{w})) \quad (3)$$

where \mathbf{Y} is the output neuron, \mathbf{X} is a set of training data, \mathbf{w} is the vector of the weight connections among the neurons and the function f is called the activation function whose form depends on the problem under consideration. For a statistical perspective on NN models, we refer to [3] and [9].

2.2 Multi-objective optimisation

In the process of optimisation, the set of experimental points $\mathbf{x}^* \in \mathbf{X}$ such that $f_k(\mathbf{x}^*) \geq f_k(\mathbf{x})$ for all $\mathbf{x} \in \mathbf{X}$ is the best composition (molecule) with respect to the objective function f_k . As mentioned in Sect. 1, in drug discovery problems molecules must satisfy different properties, introducing multiple (and possible conflicting) objective functions. This framing of the problem is referred as multi-objective optimisation and can make the search of the optimal molecules hard. Frequently in multi-objective optimisation problems, a set of experimental points \mathbf{x}^* simultaneously maximising all the objective functions does not exist. Therefore, the goal is to identify the Pareto optimal data which are the points that cannot be improved in any of the objectives without degrading at least one of the other objectives. Formally in a maximisation problem, a point $\mathbf{x}^* \in \mathbf{X}$ is a Pareto optimal solution if for every $\mathbf{x} \in \mathbf{X}$ and $k = \{1, 2, \dots, K\}$ either,

$$\forall_{k \in K} \quad f_k(\mathbf{x}) = f_k(\mathbf{x}^*) \quad (4)$$

or, there is at least one $k \in K$ such that

$$f_k(\mathbf{x}) < f_k(\mathbf{x}^*). \quad (5)$$

More details about multi-objective optimization can be found in [7].

3 Modelling the properties of MMP-12 enzyme

We investigate the performances of the model-based evolutionary approach for optimisation in a real data context regarding the study of the MMP-12 enzyme. Aim of this study is to achieve the composition of the molecules with a high potential to become a drug in the cure of certain diseases related to the enzyme¹.

3.1 Data description

The data have been originally generated and analysed by [8], and subsequently investigated by [2, 5, 12]. The whole experimental space consists of a library of 2500 molecules that have been synthesised in laboratory. In this dataset, each molecule is represented by the presence or the absence of a set of 22272 fragments, thus encoded as a binary variable where 1 represents the presence and 0 represents the absence of the specific fragment in composing the molecule. Fragments are therefore the p

¹ This research is motivated within the BLOOM project - Building Lead Optimisation Over large Molecular spaces. See <http://www.ecltech.org>

predictors and the \mathbf{X} matrix is a $N = 2500 \times p = 22272$. Several analyses have reduced the number of fragments to $p = 175$; see [5] for a detailed description. Each molecule is characterised by a set of properties:

- the pharmacological activity at the target protein, denoted by *Activity*;
- some physicochemical properties such as the capacity to dissolve in a liquid, denoted as *Solubility* and the capacity of any chemicals to produce undesirable effects denoted as *Safety*.
- some structural properties, such as the lipophilicity, denoted by *cLogP*, and the molecular weight, *MW*.

These molecular properties represent the response variables of the models Y_k , $k = 1, \dots, 5$. In order to obtain candidate molecules with suitable properties, some constraints are imposed : $Y_1=Activity > 6$, $Y_2=Solubility > -3$, $Y_3=Safety > 2.57$, $Y_4=cLogP < 3$ and $Y_5=MW < 450$. From the database made available by [8], only three molecules out of the whole library of 2500 satisfy these constraints and therefore represent the Pareto set \mathbf{x}^* of this lead optimisation problem. Based on these data, we develop our strategy to model the molecular system with the aim of optimising the associated properties. Only $n = 140$ training observations are used to avoid waste of research resources, so we proceed by generating training subsets of data of $n_m = 20$ for each generation of the process, with $m = 1, \dots, 7$. In this work we drive the evolution across generations by using predictions derived by the estimation of Lasso models and Neural Networks as described in Sect. 2.1

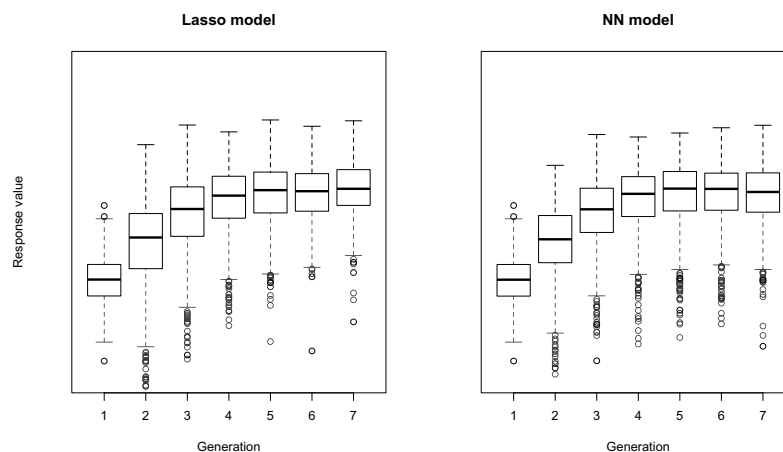
3.2 Results

Knowing the whole experimental space (complete molecular library and the associated properties) allow us to evaluate the performance of the approach in modelling and predicting the best response values with respect to changes in the initial sample data by repeating the entire process 1000 times. The performance of the procedure is therefore evaluated in its capacity to predict the response values Y_k for each molecules belonging to the test set and to identify the three molecules of the set $\mathbf{x}^* \in \mathbf{X}$, namely the Pareto set of the system. In Table 1 we present the results of the model-based evolutionary approach where the predictions driving the procedure are derived with the Lasso model and the NN model. The probability of failure, i.e. of never identifying any of the molecules belonging to the Pareto set, is very low (13% and 16.1% for Lasso and NN respectively) meaning that the estimated models are able to infer the significant properties of the molecular system.

Looking at the probability of detecting at least one of the molecules of the Pareto set, we can notice how the procedure built with both Lasso and NN models performs very good, with about 85% of times be able to uncover at least one of the best molecules out of 1000 runs (87% and 83.9% for Lasso and NN respectively). Moreover, the percentage of success increases from only one molecule to exactly all the three Pareto molecules. This shows how models perform well in modelling the rela-

Table 1 Multi-objective procedure: number of runs, out of 1000, in which the procedure uncovers the best molecules.

Predictive Models	Number of best molecules				
	0	1	2	3	At least one
Lasso	130	43	320	507	870
NN	161	59	288	492	839

**Fig. 1** Evolution through generations: boxplot of the average response values achieved in each generation for Lasso and NN model.

tions among the predictors of the system (in this study, the presence or the absence of fragments composing the molecules) and with the responses. Moreover having observed similar performances for both the Lasso and the NN models mean that the evolutionary approach built using data sampled by model predictions is guiding the researcher to detect and modelling observations of the search space where the information for the optimisation of the system properties resides. This can be seen in Fig. 1, where the evolution of the average response values achieved in each generation of the procedure is shown. From generation 1, where data are sampled at random, the information inferred by the predictive models is able to guide the search strategy to smartly identify new points, i.e. experimental points that are molecules to be tested in laboratory, with high values of experimental responses.

4 Concluding remarks

The purpose of this research is the development of a smart strategy able to address the modelling and the multi-objective optimisation of complex experimentation con-

ducting a very small number of tests. The procedure proposed is a model-based evolutionary strategy, which involves the construction and the estimation of predictive linear and non-linear models. The study of a particular molecular system for drug discovery problems shows the very good performance of the approach that we propose. Moreover, we would like to stress that we achieve these results by conducting an extremely small number of generations (7 generations) that usually is regarded too small to even approach convergence of the algorithm.

Acknowledgements The authors would like to acknowledge the fruitful collaboration with Darren Green and his Molecular Design group at GlaxoSmithKline (GSK), Medicines Research Centre, Stevenage (UK).

References

1. Baragona, R., Battaglia, F., Poli, I.: Evolutionary statistical procedures: an evolutionary computation approach to statistical procedures designs and applications. Springer, Heidelberg (2013)
2. Borrotti, M., De March, D., Slanzi, D., Poli, I.: Designing Lead Optimization of MMP-12 Inhibitors. *Comput. Math. Methods Med.*, Article ID 258627, 8 pages (2014)
3. Cheng, B., Titterington, D.M.: Neural Networks: A Review from a Statistical Perspective. *Statist. Sci.* 9(1), 2–30 (1994)
4. Ekins, S. *et al.*: Evolving molecules using multi-objective optimization: applying to ADME/Tox. *Drug Discov. Today* 15, 410–451 (2010)
5. Giovannelli, A., Slanzi, D., Khoroshiltseva, M., Poli, I.: Model-Based Lead Molecule Design in Federico Rossi; Stefano Piatto; Simona Concilio, *Advances in Artificial Life, Evolutionary Computation, and Systems Chemistry*, Springer International Publishing, 708, 103–113 (2017)
6. Li, H. *et al.*: An effective docking strategy for virtual screening based on multi-objective optimization algorithm. *BMC Bioinform.* 10(58), 1–12 (2009)
7. Lobato, F.S., Steffen, V.: Multi-objective Optimization Problem. *SpringerBriefs in Mathematics*. Springer International Publishing (2017)
8. Pickett, S.D., Green, D.V.S., Hunt, D.L., Pardoe, D.A., Hughes, I.: Automated lead optimization of MMP-12 inhibitors using a genetic algorithm. *ACS Med. Chem. Lett.* 2(1), 28–33 (2011)
9. Poli, I., Jones, R.D.: A Neural Net Model for Prediction. *J. Am. Stat. Assoc.* 89 (425), 117–121 (1994)
10. Slanzi, D., Poli, I.: Evolutionary Bayesian Network Design for High-Dimensional Experiments. *Chemometr. Intell. Lab.* **135**, 172–182 (2014)
11. Slanzi, D., De Lucrezia, D., Poli, I.: Querying Bayesian Networks to design experiments with application to IAGY serine esterase protein engineering. *Chemometr. Intell. Lab.* **149**, 28–38 (2015)
12. Slanzi D., Mameli V., Khoroshiltseva M., Poli I.: Multi-objective Optimization in High-Dimensional Molecular Systems. In: Pelillo M., Poli I., Roli A., Serra R., Slanzi D., Villani M. (eds.) *Artificial Life and Evolutionary Computation. WIVACE 2017. Communications in Computer and Information Science*, vol 830, pp.284–295. Springer, Cham (2018)
13. Soto, A.J. *et al.*: Multi-objective feature selection in QSAR using a machine learning approach. *QSAR Comb. Sci.* 28, 1509–1523 (2009)
14. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* 58(1), 267–288 (1996)

Non-crossing parametric quantile functions: an application to extreme temperatures

Il problema del crossing con funzioni quantiliche parametriche: un'applicazione alle temperature estreme

Gianluca Sottile and Paolo Frumento

Abstract Quantile regression can be used to obtain a non-parametric estimate of a conditional quantile function. The presence of quantile crossing, however, leads to an invalid distribution of the response and makes it difficult to use the fitted model for prediction. In this work, we show that crossing can be alleviated by modelling the quantile function parametrically. We then describe an algorithm for constrained optimisation that can be used to estimate parametric quantile functions with the non-crossing property. We investigate climate change by modelling the long-term trends of extreme temperatures in the Arctic Circle.

Abstract *La regressione quantilica può essere utilizzata per ottenere una stima non parametrica di una distribuzione condizionata. La presenza di crossing nei quantili stimati, tuttavia, induce una distribuzione impropria e rende difficile utilizzare il modello a scopi predittivi. Nel nostro lavoro, dimostriamo che usare una versione parametrica del modello di regressione quantilica permette di attenuare il problema del crossing. Inoltre, descriviamo un algoritmo di ottimizzazione vincolata che permette di eliminare il crossing completamente. Il metodo descritto nell'articolo viene impiegato per studiare i cambiamenti climatici e caratterizzare le tendenze a lungo termine delle temperature estreme nel Circolo polare artico.*

Key words: Parametric quantile functions, quantile regression coefficients modelling (QRCM), R package `qrcm`, estimation of extremes, climate change.

Gianluca Sottile

University of Palermo, Department of Economic, Business and Statistics, Palermo, Italy, e-mail: gianluca.sottile@unipa.it

Paolo Frumento

Karolinska Institutet, Institute of Environmental Medicine, Unit of Biostatistics, Stockholm, Sweden, e-mail: paolo.frumento@ki.se

1 Introduction

Quantile regression (QR) was first developed in [8], and is the subject of an extensive literature and an excellent monography by [9]. Quantile regression represents the natural generalisation of the concept of regression to quantile estimation. Unlike other regression methods, it is a “distribution-free” approach, as it does not require strong parametric assumptions. Although this is generally seen as an advantage, it also represents its main weakness. A well-known problem in quantile regression, which is just a consequence of its nonparametric nature, is represented by quantile crossing. Denote by \mathbf{x} the model covariates, and suppose to estimate the percentiles, corresponding to quantiles of order $\tau = \{0.01, 0.02, \dots, 0.99\}$. In principle, this enables computing a nonparametric, grid-based estimate of a conditional quantile function $Q(\tau | \mathbf{x})$. Unfortunately, at some observed values of \mathbf{x} , the estimated quantiles may not form a monotonically increasing function. While this may not hinder the interpretation of the regression coefficients, it has unpleasant consequences when the fitted model is used for prediction. For example, the estimate of the 95th percentile may be larger than that of the 99th percentile.

Crossing in parametric quantile functions

Although some authors considered various forms of parametric quantile functions, their practical use has been limited, at least in the past. A unified framework for modelling and estimation of parametric quantile functions has been described for the first time in the recent paper of [4], and in the companion R package `qrqm`. The idea is to replace a generic quantile function, say $Q(\tau | \mathbf{x})$, with a parametric model, $Q(\tau | \mathbf{x}, \boldsymbol{\theta})$. This approach has been shown to outperform ordinary quantile regression with almost no loss of information. Its advantages include smoothness, parsimony, efficiency, ease of interpretation, and a simpler computation and inference. Estimating a parametric quantile function does not automatically induce non-crossing quantiles. Imposing a parametric structure, however, alleviates the crossing problem in comparison with the nonparametric, unstructured estimator based on standard quantile regression. In this article, we introduce an algorithm for constrained optimisation that can be used to enforce the non-crossing property in parametric quantile functions.

Parametric modelling of extreme quantiles

The problem of estimating extreme quantile dates back to [5] and [6], among others. Approaches based on quantile regression have been described by different authors, including [1, 14, 3]. In some of these works, a grid of “intermediate” quantiles is used to extrapolate the extremes. If the quantile function is modelled parametrically, extreme quantiles are just functions of the model parameters, and can be directly extrapolated from the fitted model. The imposed parametric structure automatically transfers information from intermediate to extreme quantiles, which can significantly reduce variability in comparison with standard quantile regression estimators. However, because crossing occurs mainly in the tails of the distribution, it is not plausible to estimate extreme quantiles without imposing monotonicity constraints.

Outline of the paper

The paper is structured as follows. In Section 2 we review the existing literature on parametric quantile functions. In Section 3 we revisit the quantile crossing problem in a parametric world. In Section 4 we present a method to estimate non-crossing quantile functions. In Section 5 we analyse the data that motivated this project.

2 Parametric quantile functions

We denote by $Q(\tau | \mathbf{x})$ the quantile function of a response variable Y , conditional on a q -dimensional vector \mathbf{x} of covariates. A standard representation of $Q(\tau | \mathbf{x})$ is the linear quantile function $Q(\tau | \mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}(\tau)$ in which $\boldsymbol{\beta}(\tau)$ is a vector of unknown regression coefficients. In standard quantile regression (QR; e.g., [8, 9]), $\boldsymbol{\beta}(\tau)$ is estimated as the minimiser of

$$L(\boldsymbol{\beta}(\tau)) = \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}(\tau)) \quad (1)$$

where $\rho_\tau(u) = (\tau - I(u \leq 0))u$. Because quantiles of different order are estimated individually, standard QR is thought of as a nonparametric method, the only “parametric” assumption being the functional form of the linear predictor.

In their recent paper, [4] suggested replacing $Q(\tau | \mathbf{x})$ by a parametric quantile function $Q(\tau | \mathbf{x}, \boldsymbol{\theta})$, and reformulated the model as $Q(\tau | \mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^T \boldsymbol{\beta}(\tau | \boldsymbol{\theta})$. Their approach is referred to as *quantile regression coefficients modelling* (QRCM), and is currently implemented in two R packages, `qrcm` and `qrcmNP`. A convenient parametrisation of this model is obtained when $\boldsymbol{\beta}(\tau | \boldsymbol{\theta})$ is expressed by a linear combination of known “basis” functions, $\mathbf{b}(\tau) = [b_1(\tau), \dots, b_k(\tau)]^T$, such that $\boldsymbol{\beta}(\tau | \boldsymbol{\theta}) = \boldsymbol{\theta} \mathbf{b}(\tau)$. In this model, $\boldsymbol{\theta}$ is a $q \times k$ matrix, and the quantile function can be rewritten as $Q(\tau | \mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^T \boldsymbol{\theta} \mathbf{b}(\tau)$. For instance, if $\mathbf{b}(\tau) = [1, \tau, \tau^{1/2}]^T$, regression coefficients are given by $\beta_j(\tau | \boldsymbol{\theta}) = \theta_{0j} + \theta_{1j}\tau + \theta_{2j}\tau^{1/2}$, $j = 1, \dots, q$. This type of parametrisation substantially simplifies computation and inference, without constituting a real limitation in terms of model flexibility. The unknown parameter vector $\boldsymbol{\theta}$ is estimated as the minimiser of

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \int_0^1 \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}(\tau | \boldsymbol{\theta})) d\tau \quad (2)$$

which is the integral, with respect to τ , of the loss of ordinary quantile regression defined in (1).

3 Crossing in parametric quantile functions

Assume to estimate a quantile function $Q(\tau | \mathbf{x}, \boldsymbol{\theta})$, and denote by $Q'(\tau | \mathbf{x}, \boldsymbol{\theta})$ its first derivative. In the standard linear quantile regression model, $Q(\tau | \mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^T \boldsymbol{\beta}(\tau | \boldsymbol{\theta})$ and $Q'(\tau | \mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^T \boldsymbol{\beta}'(\tau | \boldsymbol{\theta})$, which simplify to $Q(\tau | \mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^T \boldsymbol{\theta} \mathbf{b}(\tau)$ and $Q'(\tau | \mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^T \boldsymbol{\theta} \mathbf{b}'(\tau)$. Denote by $\hat{\boldsymbol{\theta}}$ the minimiser of the loss function defined in equation (2). Quantile crossing occurs if the set $\{\tau : Q'(\tau | \mathbf{x}, \hat{\boldsymbol{\theta}}) < 0\}$ is non-empty. Crossing may be due to a variety of reasons, some of which are illustrated by the following examples, with the graphical support of Figure 1.

Example (a). Misspecified model with empty feasible region. Consider the quantile function defined by $Q(\tau | x, \theta) = \theta \tau x$. As shown in Figure 1a, all regression lines cross at $x = 0$, where the model is assumed to be degenerated. If x only takes either positive or negative values, this model is guaranteed to be non-crossing. Otherwise, crossing occurs at all values of θ and cannot be avoided.

Example (b). Crossing caused by outliers. Define $Q(\tau | x, \boldsymbol{\theta}) = \beta_0(\tau | \boldsymbol{\theta}) + \beta_1(\tau | \boldsymbol{\theta})x$ with $\beta_0(\tau | \boldsymbol{\theta}) = \theta_{00} + \theta_{01}\tau$ and $\beta_1(\tau | \boldsymbol{\theta}) = \theta_{10} + \theta_{11}\tau$, and assume that x lies in the $(0, 1)$ interval. A simple condition for monotonicity is that $(\hat{\theta}_{01}, \hat{\theta}_{11}) \geq 0$. However, as shown in Figure 1b, the estimated regression lines may cross in correspondence of outlying values.

Example (c). Non-monotone coefficient. Define $Q(\tau | x, \boldsymbol{\theta}) = \theta_0 \tau - \theta_1(\tau - 0.75)^2 x$, and assume $\boldsymbol{\theta} > \mathbf{0}$ and $x \geq 0$. The regression model is illustrated in Figure 1c. The coefficient associated with x is assumed to be a non-monotone function, $\beta_1(\tau | \boldsymbol{\theta}) = -\theta_1(\tau - 0.75)^2$, which may obviously induce crossing. A non-crossing quantile function can be obtained by imposing $2\hat{\theta}_0 \geq \hat{\theta}_1 \max_i(x_i)$.

Example (d). Crossing in a flexible model. In most situations, the true model is not known and the coefficients are described by splines, polynomials, or other types of smooth flexible functions. Consider, for example, the data illustrated in Figure 1d, and suppose to fit a quantile function $Q(\tau | x, \boldsymbol{\theta}) = \beta_0(\tau | \boldsymbol{\theta}) + \beta_1(\tau | \boldsymbol{\theta})x + \beta_2(\tau | \boldsymbol{\theta})x^2 + \beta_3(\tau | \boldsymbol{\theta})x^3$, in which $\beta_j(\tau | \boldsymbol{\theta}) = \theta_{j0} + \theta_{j1}\tau + \theta_{j2}\tau^2 + \theta_{j3}\tau^3$, $j = 0, \dots, 3$. In this example, quantiles are cubic functions of x , and regression coefficients are cubic functions of τ . Crossing at extreme quantiles arises from the combination between a very flexible parametric structure and a relatively small sample size.

Situations like those described in Example (a), in which the feasible region is empty, must be considered pathological and may only occur if the model is severely misspecified and is subject to restrictive assumptions. Example (c) suggests that crossing can be attenuated or completely eliminated by a suitable parametrisation of the quantile function. For example, a general form of non-crossing quantile function is obtained when $\mathbf{x} \geq \mathbf{0}$ and $\boldsymbol{\beta}'(\tau | \boldsymbol{\theta}) > \mathbf{0}$, i.e., all coefficients are monotonically increasing functions of τ . This model can be used to build growth charts (e.g., [10]). A special case of it is the constant-slope model, in which all coefficients apart from the intercept are assumed to not depend on τ .

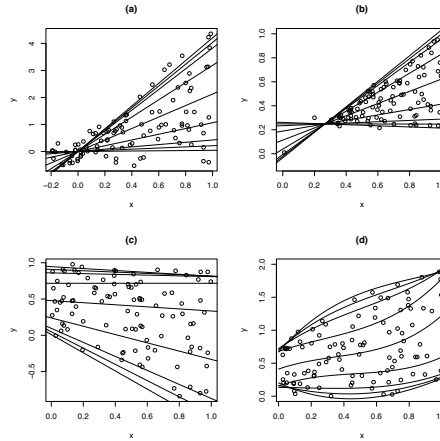


Fig. 1 Simulated data with superimposed parametric models suffering from quantile crossing. Lines represent estimated quantiles of order $\tau = \{0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99\}$. Figures **a-d** correspond to examples **a-d** in the text.

4 Parametric estimation of non-crossing quantile functions

We describe a numerical method to estimate a parametric quantile function with the non-crossing property. The problem to be solved is the following:

$$\begin{aligned} & \min L(\boldsymbol{\theta}) \\ \text{s.t. } & \int_0^1 |\min\{0, Q'(\tau | \mathbf{x}_i, \boldsymbol{\theta})\}| d\tau = 0, \quad i = 1, \dots, n. \end{aligned} \quad (3)$$

where $L(\boldsymbol{\theta})$ is the loss function defined in equation (2). Each constraint is a function of $\boldsymbol{\theta}$, and is equal to zero if no crossing occurs, and a positive quantity otherwise. To solve problem (3), we suggest using a penalised approach as in [12] and [13], and estimate $\boldsymbol{\theta}$ as the minimiser of

$$\begin{aligned} L_\lambda(\boldsymbol{\theta}) &= L(\boldsymbol{\theta}) + \lambda P(\boldsymbol{\theta}) \\ &= \sum_{i=1}^n \int_0^1 \rho_\tau(y_i - Q(\tau | \mathbf{x}_i, \boldsymbol{\theta})) d\tau + \lambda \sum_{i=1}^n \int_0^1 |\min\{0, Q'(\tau | \mathbf{x}_i, \boldsymbol{\theta})\}| d\tau, \end{aligned} \quad (4)$$

where λ is a positive tuning constant, and $P(\boldsymbol{\theta})$ is a penalty term computed as the sum of all constraints. The value of $P(\boldsymbol{\theta})$ reflects both the sign and the absolute size of $Q'(\tau | \mathbf{x}_i, \boldsymbol{\theta})$, and is zero at all values of $\boldsymbol{\theta}$ that generate a non-crossing quantile function. The tuning parameter λ defines the balance between the two ingredients of $L_\lambda(\boldsymbol{\theta})$, namely the unpenalised loss function, $L(\boldsymbol{\theta})$, and the penalty term, $P(\boldsymbol{\theta})$. If λ is too small, the penalisation has no effect and the constraints are ignored. As λ increases, the relative weight of $P(\boldsymbol{\theta})$ becomes larger, and $\hat{\boldsymbol{\theta}}$ is pushed towards the feasible region. However, if λ is unnecessarily large, convexity of $L_\lambda(\boldsymbol{\theta})$ is no

longer guaranteed, because $P(\boldsymbol{\theta})$, unlike $L(\boldsymbol{\theta})$, is not generally a convex function. The minimisation algorithm is initialised with a relatively small value of λ , which is progressively increased until convergence.

5 Long-term trends of extreme temperatures

In the last decades, global warming has been the subject of an increasingly large number of articles and scientific reports (e.g., [2]), and is now considered one of the most important challenges of mankind in the coming century. The analysis of climatological data is often based on relative simple indicators, such as the frequency of days in which the minimum (maximum) temperature is below (above) a certain threshold of interest. Any summary measure, however, can be expressed as a function of the conditional quantiles. In particular, extreme events correspond to very low or high quantiles of the distribution of a meteorological variable.

In this work, we considered meteorological data from ECAD (European Climate Assessment and Dataset). The data are described in [7] and can be downloaded from the ECAD website (<https://www.ecad.eu/>). The datasets contain daily information on temperature, precipitations, and cloud cover from over 4,000 meteorological stations across Europe. Most time series start in the early 1900s and can be utilised to investigate long-term trends.

We considered the minimum and the maximum daily temperature, particularly focusing on the extreme quantiles, corresponding to cold/heat waves. We formulated the following regression model, to be applied to both responses separately:

$$Q(\tau | t, s) = \beta_0(\tau) + \mathbf{g}_t(\mathbf{t})^T \boldsymbol{\beta}_t(\tau) + \mathbf{g}_s(\mathbf{s})^T \boldsymbol{\beta}_s(\tau) + (\mathbf{g}_t(\mathbf{t}) \otimes \mathbf{g}_s(\mathbf{s}))^T \boldsymbol{\beta}_{t,s}(\tau). \quad (5)$$

In the above regression equation, t is a progressive count (1, 2, ...), expressed in days, while $s = (t \bmod 365.2422)$ counts the days within a solar year (365.2422 days). The long-term trend corresponds to the “effect” of t , and is modelled by the basis of a restricted natural cubic spline, $\mathbf{g}_t(\mathbf{t})$, with one internal knot every 20 years. Seasonal variations are described by a periodic spline, $\mathbf{g}_s(\mathbf{s})$, with a period of one solar year and one internal knot every 2 solar months. Additionally, the regression equation includes the tensor product $\mathbf{g}_t(\mathbf{t}) \otimes \mathbf{g}_s(\mathbf{s})$ that defines an interaction term.

Crossing represented a very common problem in our analysis. It especially affected the tails of the distribution, and often prevented using the fitted model to extrapolate future trends. Consider, for example, station n. 78, that is located at 26 meters above sea level in Kandalaksha, Russia, with coordinates (N 67° 09' 00", E 32° 21' 00"). Data for this station are available from 1912 to 2018. As shown in Table 1, standard quantile regression (QR) estimators suffered from severe crossing. The problem was significantly alleviated by using a parametric model (QRCM) in which all coefficients in model (5) were described by a 5th-degree polynomial function of τ . Finally, a constrained version of this model (QRCM_c) allowed to eliminate crossing entirely.

Table 1 Indicators of crossing using different methods for quantile regression to analyse the minimum and maximum temperature at the Kandalaksha station, Russia, between 1912 and 2018. In the table, QR denotes standard quantile regression, while QRCM and QRCM_c indicate unconstrained and constrained parametric estimators, respectively. P_{cross} is computed as the proportion of observations for which the estimated quantile function was non-monotone; and L_{cross} is the average length of the crossing region on the τ scale.

		QR	QRCM	QRCM _c
Minimum temperatures	$100 \times P_{\text{cross}}$	45.03	2.94	0.00
	$100 \times L_{\text{cross}}$	2.69	0.23	0.00
Maximum temperatures	$100 \times P_{\text{cross}}$	34.56	0.32	0.00
	$100 \times L_{\text{cross}}$	2.56	0.03	0.00

Estimators based on constrained minimisation are illustrated in Figures 4.2 and 4.3, where we report model-based quantiles of the minimum and maximum temperature. We use a variety of predictions to describe the long-term and seasonal trends. While the seasonal trends appear unchanged, the long-term trends show a cooling between the 1960s and the early 1990s, [11], followed by a warming effect. Would the current trends continue in the future, in twenty years we may expect an additional warming effect well above 1°C in both winter and summer.

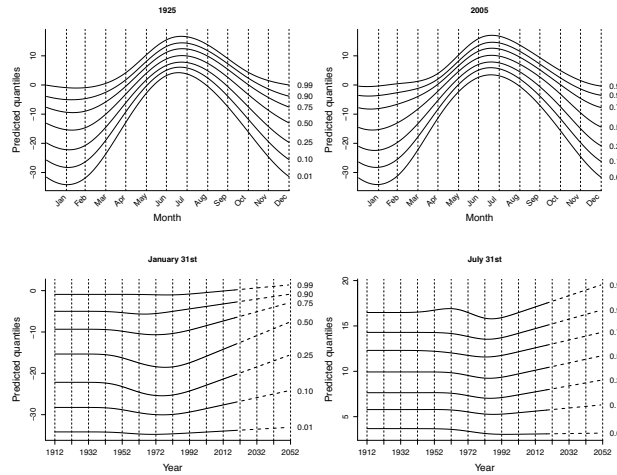


Fig. 4.2 Estimated quantiles of order $\tau = \{0.01, 0.10, 0.25, 0.50, 0.75, 0.90, 0.99\}$ of the minimum daily temperature in Kandalaksha station, Russia. The top panels compare the seasonal trend in 1925 and that in 2005, while the bottom panels illustrate the long-term trends on two selected days, Jan 31st and July 31st. Dashed lines indicate extrapolation.

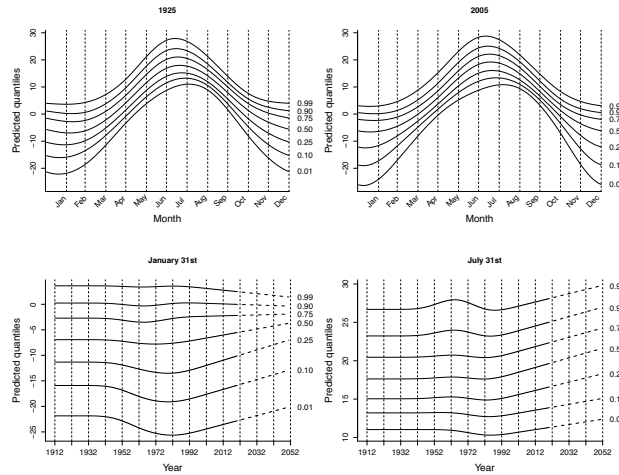


Fig. 4.3 Estimated quantiles of the maximum temperature in Kandalaksha station, Russia.

References

1. Chernozhukov, V: Extremal quantile regression. *The Annals of Statistics*, **33**, 806–839 (2005).
2. Cox, PM, Betts, RA, Jones, CD, Spall, SA, Totterdell, IJ: Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature*, **408**(6809), 184(4) (2000).
3. De Valk, C: Approximation of high quantiles from intermediate quantiles. *Extremes*, **19**(4), 661–686 (2016a).
4. Frumento, P, Bottai, M: Parametric modeling of quantile regression coefficient functions. *Biometrics*, **72**(1), 74–84 (2016).
5. Gnedenko, BV: Sur la distribution limite du term maximum d’une sèrie aleatoire. *Annals of Mathematics*, **44**(3), 423–453 (1943).
6. Hill, BM: A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, **3**(5), 1163–1174 (1975).
7. Klein Tank AMG and Coauthors: Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, **22**, 1441–1453 (2002).
8. Koenker, R: *Quantile regression*. Econometric Society Monograph Series, Cambridge: Cambridge University Press. (2005)
9. Koenker, R, Bassett GJr: Regression Quantiles. *Econometrica*, **46**, 33–50 (1978).
10. Muggeo, VMR, Sciandra, M, Tomasello, A, and Calvo, S: Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. *Environmental and Ecological Statistics*, **20**(4), 519–531 (2013).
11. Peterson, TC, Connolley, WM, Fleck, J: The myyth of the 1970s global cooling scientific consensus. *American Meteorological Society*, **89**(9), 1325–1337 (2008).
12. Richardson, JT, Palmer, MR, Liepins, G, Hilliard, M: Some guidelines for genetic algorithms with penalty functions. *Proceedings of the Third International Conference on Genetic Algorithms*, 191–197 (1989).
13. Siedlecki, W, Sklansky, J: Constrained genetic optimization via dynamic reward-penalty balancing and its use in pattern recognition. *Proceedings of the Third International Conference on Genetic Algorithms*, 141–150 (1989).
14. Wang, H, Li, D, He, X: Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association*, **107**, 1453–1464 (2012).

A new tuning parameter selector in lasso regression

Un nuovo criterio di selezione per il parametro di penalizzazione nella regressione lasso

Gianluca Sottile and Vito MR Muggeo

Abstract Penalized regression models are popularly used in high-dimensional data analysis to carry out variable selection and model fitting simultaneously. Whereas success has been widely reported in literature, their performance largely depend on the tuning parameter that balances the trade-off between model fitting and sparsity. In this work we introduce a new tuning parameter selection criterion based on the maximization of the signal-to-noise ratio. To prove its effectiveness we applied it to a real data on prostate cancer disease.

Abstract *I modelli di regressione penalizzata vengono comunemente utilizzati nell'analisi di dati ad alta dimensionalità per condurre simultaneamente la selezione delle variabili più informative e il buon adattamento del modello. Mentre il successo di tali tecniche è stato ampiamente riportato in letteratura, le loro prestazioni dipendono in gran parte dai parametri di penalizzazione. In questo lavoro introduciamo un nuovo criterio di selezione del parametro di penalizzazione basato sull'idea della massimizzazione del rapporto segnale-rumore. Per provare la sua efficacia abbiamo applicato tale criterio ad un dataset reale sul cancro alla prostata.*

Key words: Least absolute shrinkage and selection operator (lasso), Model selection, Variable selection, Penalized likelihood, Signal-to-noise ratio, Clinical data.

Gianluca Sottile

University of Palermo, Department of Economic, Business and Statistics, Palermo, Italy, e-mail: gianluca.sottile@unipa.it

Vito MR Muggeo

University of Palermo, Department of Economic, Business and Statistics, Palermo, Italy, e-mail: vito.muggeo@unipa.it

1 Introduction

In the context of high-dimensional data, typically only a small number of variables are truly informative whilst most of them are redundant. Selecting the appropriate variables is a crucial step of the data analysis process. In fact underfitted models excluding truly informative covariates may lead to severe bias of the estimators. On the other hand, overfitting may hinder interpretation and cause large standard errors [1]. Penalized regression methods have gained popularity as a tool to perform variable selection. These methods operate maximizing the penalized likelihood function

$$\frac{1}{n} \ell(\boldsymbol{\beta}) - \sum_{j=1}^d p_{\lambda}(|\beta_j|) \quad (1)$$

with respect to $\boldsymbol{\beta} \in \mathbb{R}^d$, where n is the sample size, $\ell(\boldsymbol{\beta})$ is the working log-likelihood function, d is the total number of covariates, and $p_{\lambda}(\cdot)$ is a penalty function that penalizes against model complexity and the size of the estimated coefficients. Increasing the amount of the regularization, defined by the positive tuning parameter λ , increases the number of estimated non null coefficients thus performing automatic variable selection. For most penalty objective like 1 efficient algorithms [16] already exist making possible to do variable selection in high dimensions.

The performance of the estimated model heavily depends on the choice of the tuning parameter. Among the different penalized procedures, the least absolute shrinkage and selection operator (lasso, [2]) appears to be the most widely utilized approach. In particular, in lasso regression, the penalty function reduces to $p_{\lambda}(\cdot) = \lambda \sum_{j=1}^d |\beta_j|$.

Since the choice of the tuning parameter balances the trade-off between model fit and model sparsity, the selection of an appropriate value is fundamental in penalized regressions. Several classical model selection procedures have been heuristically applied as selectors of this parameter including information criteria, e.g., simple and generalized cross-validation (respectively CV and GCV, [3]), Akaike information criterion (AIC, [4]), Bayesian information criterion (BIC, [5]) and its extended version (EBIC, [6]), the more recent Generalized information criterion (GIC, [7]) and stability selection ([8]). The statistical properties of these criteria have been widely studied in the context of classical regression. Unfortunately, no method appears to perform systematically better than others, and many authors proposed to select the tuning parameters through the data-driven approach, i.e., the k -fold CV, which is also the default option in several R packages.

In this work we propose a new criterion to select the tuning parameter in lasso regression. The criterion is quite simple to compute and relies on maximization of the signal-to-noise ratio (SNR). We show that our tuning parameter selector enables to identify the true model consistently when the true model is among a set of candidate models.

The rest of the paper is organized as follows. Section 2 presents the proposed model selector. Application to real examples are presented in Section 3 to illustrate the use in practice. Section 4 provides the final discussions.

2 New selection criterion

2.1 Penalized linear regression framework

Consider data $(\mathbf{x}_1^T, y_1), \dots, (\mathbf{x}_n^T, y_n)$ where y_1, \dots, y_n are independent given \mathbf{X} . y_i is the response from the i -th subject, and \mathbf{x}_i is the associated d -dimensional covariate vector. Let consider the gaussian model $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ is a $n \times 1$ unknown mean vector and the entries of the $n \times 1$ error vector $\boldsymbol{\varepsilon}$ are independent and identically distributed (iid) with mean 0 and variance σ^2 . The mean vector is estimated by $\hat{\boldsymbol{\mu}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{\beta}}_\lambda$ is the estimator that minimize the penalized least squares function

$$\frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^d |\beta_j| \quad (2)$$

with respect to $\boldsymbol{\beta} \in \mathbb{R}^d$.

2.2 Recap of Theoretical assumptions of the lasso framework

We assume that the set of candidate models contains the unique true model, and that the number of parameters in the full model is finite. Being β_j^0 the true vector of coefficients and $S_0 := \{j : \beta_j^0 \neq 0\}$ the true active set, any selection criteria deliver an estimator \hat{S} of S_0 such that $\hat{S} = S_0$ with large probability. However, it is clear that very small coefficients $|\beta_j^0|$ can not be detected, hence a first condition to obtain perfect recovery is the so-called “*beta-min condition*”, i.e., $\min_{j \in S_0} |\beta_j^0| \geq c \cdot \sqrt{2\phi \log d}$ where c is an unknown numerical constant (which would have to exceed one) and ϕ is the dispersion parameter associated to the log-likelihood function of the exponential family. A second condition is about the degree of sparsity, i.e., let be d_0 the cardinality of S_0 then the true number of nonzero coefficients in the model have to obey to $d_0 \leq n/(2 \log d)$.

This two conditions belongs to the so-called *linear sparsity* conditions [9, 10, 11, 12], that allow to describe the performance of the lasso in practical settings with moderately large dimensions and reasonable values of the degree of sparsity. Moreover, for consistent variable selection a sufficient and necessary condition, is the “*irrepresentable condition*” or “*restricted eigenvalue condition*” [13, 14, 15], that is a condition based on the structure of the design matrix. Usually, if all conditions are satisfied, each traditional criterion is able to discover the true nonzero coefficients.

2.3 Recap of previous proposals

Previous proposal as discussed in Section 1 are AIC, BIC, EBIC, GCV, GIC and 5-fold CV, and their formulations are given as follows,

$$\begin{aligned}
\text{AIC} &= \log(\hat{\sigma}_\lambda^2) + 2 d_\lambda n^{-1} \\
\text{BIC} &= \log(\hat{\sigma}_\lambda^2) + \log n d_\lambda n^{-1} \\
\text{EBIC} &= \log(\hat{\sigma}_\lambda^2) + (\log n + 2\gamma \log d) d_\lambda n^{-1} \\
\text{GCV} &= \hat{\sigma}_\lambda^2 / (1 - d_\lambda n^{-1})^2 \\
\text{GIC} &= \log(\hat{\sigma}_\lambda^2) + c_n \log d d_\lambda n^{-1} \\
\text{5-fold CV} &= \sum_{s=1}^5 \sum_{(y_k, \mathbf{x}_k) \in T^{-s}} \left(y_k - \mathbf{x}_k^T \hat{\boldsymbol{\beta}}_\lambda^{(s)} \right)^2
\end{aligned}$$

where $\hat{\sigma}^2$ is the estimated dispersion parameter that could be estimated as $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\|_2^2 / (n - d_\lambda)$, γ is a non-negative parameter and c_n is a parameter which depends on n .

2.4 The proposed criterion

In literature many definitions of the SNR have been proposed and the most used by many authors is given by $\|\boldsymbol{\mu}\|_2^2 / \phi$ that in linear regression models could be easily written as $\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} / \phi$. However, in the special case in which the design matrix is orthonormal, i.e., $\mathbf{X}^T \mathbf{X} = \mathbf{I}$, an ℓ_1 -view of it is given by $\|\boldsymbol{\beta}\|_1 / \phi^{1/2}$. Since we could estimate as many SNRs as the length of the λ values, we suggest to select λ as the maximizer of the weighted signal-to-noise ratio (WSNR)

$$\arg \max_{\lambda} w_\lambda \frac{\|\hat{\boldsymbol{\beta}}_\lambda\|_1}{\hat{\sigma}_\lambda}, \quad (3)$$

where $w_\lambda = d_\lambda^{-1}$ is the model degrees of freedom, namely the cardinality of the active set, i.e., $|S_\lambda| = |\{j : \hat{\beta}_{\lambda,j} \neq 0\}|$ and σ_λ is the square root of the dispersion parameter that could be fixed or estimated through the data.

3 Prostate Cancer data set

The real data used is the well-known Prostate Cancer data from a study on prostate cancer [17], measuring the correlation between the level of a prostate-specific antigen and some covariates. The covariates are $x_1 = \text{lccavol}$ (log-cancer volume), $x_2 = \text{lweight}$ (log-prostate weight), $x_3 = \text{age}$ (age of patient), $x_4 = \text{lbhpl}$ (log-amount of benign hyperplasia), $x_5 = \text{svi}$ (seminal vesicle invasion), $x_6 = \text{lcp}$ (log-capsular

penetration), $x_7 = \text{gleason}$ (Gleason Score), $x_8 = \text{pgg45}$ (percent of Gleason scores 4 or 5) and response variable $y = \text{lpsa}$ (log-psa). The data set consists of $n = 97$ observations and $p = 8$ covariates. To assess the performance of all regularization parameter selectors, the same procedure as for the riboflavin data set is applied ($n_{\text{ts}} = 73, n_{\text{vs}} = 24$).

Table 1 Tuning parameter selection of the Prostate Cancer data set. Different criteria are used ASNR, AIC, BIC, EBIC, GCV, GIC and 5-fold CV. The number of nonzero coefficients, the tuning parameter selected (λ^*) and the prediction error are reported. An intercept is added to the model.

	# of Nonzero Coefficients	λ^*	Prediction Error
WSNR	3	0.1541	0.3994
AIC	4	0.0608	0.3990
BIC	4	0.0608	0.3990
EBIC	3	0.1541	0.3994
GCV	4	0.0608	0.3990
GIC	4	0.0608	0.3990
CV	6	0.0289	0.4027

Table 1 reports the number of nonzero coefficients, the tuning parameter selected and the mean prediction error measure for each criterion. It is possible to see that WSNR and EBIC selected the same tuning parameter identifying, as already known in literature, the three non-zero covariates *lcavol*, *lweight* and *svi*. Figure 1 shows the graphical representation of each criterion.

4 Discussion

In the context of variable selection, we propose a new information criterion to choose regularization parameter. Furthermore, we study the theoretical properties of ASNR. If we believe that the true model is contained in a set of candidate models with the generalized linear model structure, then our selector identifies the true model consistently, while the other criteria tend to overfit. Simulation studies and empirical examples support the performance of the selection criteria.

Even if the theoretical property of WSNR is not yet formally proven, the empirical results suggest the potential of it. Moreover, our proposal can be extended to generalized linear models, e.g., Poisson and logistic regression. Application in very high-dimensional settings ($n \ll p$) that are today one of the most challenging concerns, represents a noteworthy application to be investigated.

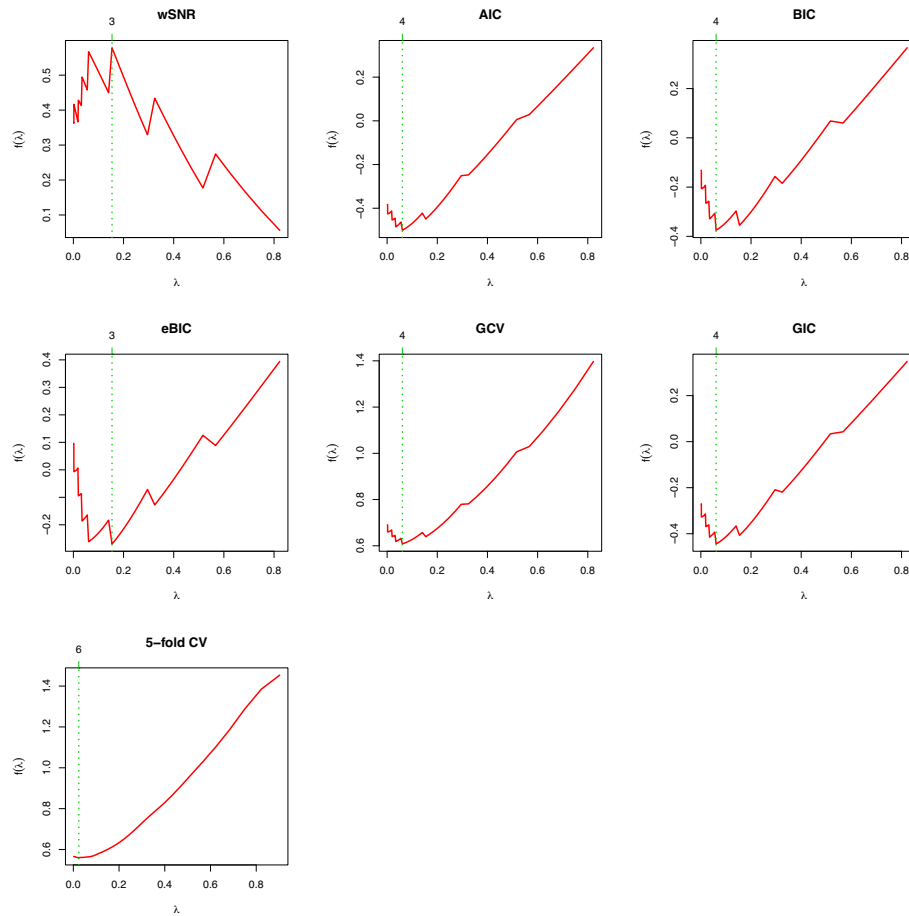


Fig. 1 Graphical curves of different tuning parameter criteria applied to the Prostate Cancer data set. The reported criteria are ASNR, AIC, BIC, EBIC, GCV, GIC and 5-fold CV. The number of non null estimated coefficients is reported on the top of each panel.

References

1. Fan, J., and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, 96, 1348–1360.
2. Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288.
3. Craven, P., and Wahba, G. (1979), “Smoothing noisy data with spline functions,” *Numerische Mathematik*, 31, 377–403.
4. Akaike, H. (1974), “A new look at the statistical model identification,” *IEEE Trans. on Automatic Control*, 19, 716–723.
5. Schwarz, G. (1978), “Estimating the dimension of a model,” *The Annals of Statistics*, 6, 461–464.

6. Chen, J., and Chen, Z. (2008), “Extended Bayesian information criteria for model selection with large model spaces,” *Biometrika*, 95, 759–771.
7. Zhang, Y., Li, R., and Tsai, C.-L. (2010), “Regularization parameter selections via generalized information criterion,” *Journal of the American Statistical Association*, 105, 312–323.
8. Meinshausen, N., and Bühlmann, P. (2010), “Stability selection,” *Journal of the Royal Statistical Society: Series B*, 72, 417–473.
9. Wainwright, M. J. (2009a), “Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting,” *IEEE Transactions on Information Theory*, 55, 5728–5741.
10. — (2009b), “Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1 -constrained quadratic programming (Lasso),” *IEEE transactions on information theory*, 55, 2183–2202.
11. Reeves, G., and Gastpar, M. C. (2013), “Approximate sparsity pattern recovery: Information-theoretic lower bounds,” *IEEE Transactions on Information Theory*, 59, 3451–3465.
12. Su, W., Bogdan, M., and Candes, E. (2015), “False discoveries occur early on the lasso path,” *arXiv preprint arXiv:1511.01957*.
13. Zou, H. (2006), “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, 101, 1418–1429.
14. Zhao, P., and Yu, B. (2006), “On model selection consistency of Lasso,” *Journal of Machine learning research*, 7, 2541–2563.
15. Bühlmann, P., and Van De Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.
16. Friedman, J., Hastie, T., and Tibshirani, R. (2010), “Regularization Paths for Generalized Linear Models via Coordinate Descent,” *Journal of statistical software*, 33, 1–22.
17. Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. (1989), “Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients.” *The Journal of Urology*, 141, 1076–1083.

Similarity patterns, topological information and credit scoring models

Strutture di similarità, informazioni topologiche e modelli di credit scoring

Alessandro Spelta, Branka Hadji-Misheva and Paolo Giudici

Abstract In this paper we show that credit risk accuracy of peer-to-peer platforms could be improved by leveraging topological information embedded into similarity networks derived from borrowers balance-sheet features. Relevant patterns of similarities describing institutions importance and community structures are extracted from the networks and employed as additional explanatory variables for improving the performance of different classes of scoring models. Results suggest that this approach is effective for discriminating between default and sound institutions and the proposed methodology can constitute a new instrument in both policy-makers and practitioners toolboxes.

Abstract In questo lavoro mostriamo che l'accuratezza nella misurazione del rischio di credito delle piattaforme peer-to-peer potrebbe essere migliorata sfruttando le informazioni topologiche contenute nelle reti di similarità derivate dalle caratteristiche di bilancio dei richiedenti prestiti. Strutture di similarità che descrivono l'importanza delle istituzioni e le strutture dei clusters sono estratte dalle reti e impiegate come variabili esplicative aggiuntive per migliorare le prestazioni di diverse classi di modelli di rating. I risultati suggeriscono che questo approccio efficace per discriminare tra istituzioni a rischio e istituzioni solide e la metodologia proposta può costituire un nuovo strumento sia per i policy-makers che per i professionisti.

Key words: Peer-to-Peer Lending Credit Scoring Models, Networks

Alessandro Spelta
University of Pavia, via San Felice 5, e-mail: alessandro.spelta@unipv.it

Branka Hadji-Misheva
University of Pavia, via San Felice 5 e-mail: branka.hadjimisheva01@universitadipavia.it

Paolo Giudici
University of Pavia, via San Felice 5 e-mail: paolo.giudici@unipv.it

1. Introduction

Fintech services, such as peer-to-peer lending platforms, are becoming part of the everyday life of millions of individuals. Such new technologies can increase financial inclusion, but they can come out at the cost of an increase credit risk. All amplified by systemic risk, due to the high interconnectdness of fintech platforms, which increases contagion. Despite the fact that both classic banks and peer-to-peer platforms rely on credit scoring models for estimating credit risk, incentives for having the best model are different since peer-to-peer lending platforms do not internalize credit risk that is fully borne by the lender. Against this background, fintech risk management becomes a central point of interest for regulators and supervisors, to protect consumers and preserve financial stability. In this paper we show that topological information embedded into similarity networks can be exploited to increase the predictive performance of some credit scoring models (see Khandani et al. (2010); Khashman (2011); Lessmann et al. (2015); Abellán and Castellano (2017) to cite few). Similarity patterns between companies' features can be extracted from a distance matrix and they can reveal how credit risk is related to the topology of the network. To summarize such topological information we compute centrality measures and community detection on two type of data sets.

Various measures of centrality have been proposed in network theory such as the count of neighbors of a node has, i.e. the degree centrality, which is a local centrality measure, or measures based on the spectral properties of the graph (see Perra and Fortunato (2008)). Spectral centrality measures include the eigenvector centrality (Bonacich (2007)), Katzs centrality (Katz (1953)), PageRank (Brin and Page (1998)), hub and authority centralities (Kleinberg (1999)). These measures are feedback, also know as global, centrality measures and provide information on the position of each node relative to all other nodes. For our purposes we employ both families of centrality measures. In particular, for each node we compute the degree and strength centrality together with the PagePank centrality.

In general, centrality measures rank vertices according to their systemic importance without paying attention to whether the network is characterized by a community structure. On the contrary, several studies have analyzed the empirical characteristics of different networks have found the presence of sets of nodes usually defined as very dense subgraphs, with few connections between them, as a result of similar patterns at the micro-level (see Pecora et al. (2016); Spelta et al. (2018)). The Louvain Method for community detection is a method to extract communities from large networks created by Blondel et al. (2008).

The forecasting gain obtained by the inclusion of these variables has measured applying concepts derived from non-parametric statistics. Beside standard performance measure such as ROC, precision recall and accuracy we also compute the Net Reclassification Improvement derived from the inclusion of topological measures. Results reveal the usefulness of the proposed methodology to build an early-warning signal suitable for both policy-makers and practitioners.

2. Results

This Section is devoted to show the results of the analysis. First we report, for both data sets, the Minimal Spanning Tree representation of the similarity network obtained from companies' feature distances. We report two types of network visualization, the first one shows nodes according to their financial soundness, red nodes represent defaulted institutions while green nodes represent sound companies, see Figure 1. Notice how, for both data set, defaulted institutions occupy precise portion of the network, namely, such companies belong to the leafs of the tree and form clusters. This, in other words, suggests those companies share particular feature together. To highlight the retrieved clusters we also report plots in which nodes are colored according to the community they belong.

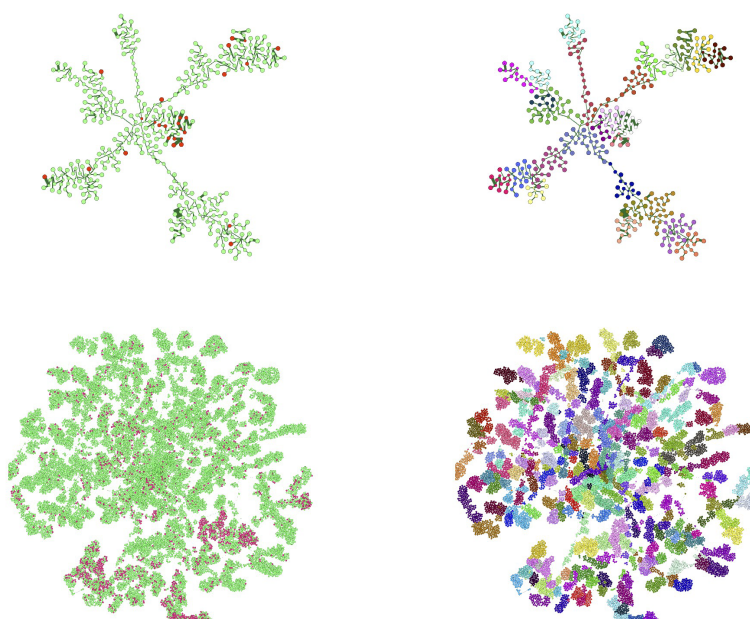


Figure 1: Minimal spanning tree representation of the borrowing companies network. The tree has been obtained by using the standardized Euclidean distance between institutions features and the Kruskal algorithm. In the left panels, nodes are colored according to their financial soundness, red nodes represent defaulted institutions while green nodes are associated with sound companies. The right panels show the same networks but nodes are colored according to the community they belong. Notice how defaulted institutions strongly occupy certain specific communities not being equally distributed among the network. Moreover, the upper panels refer to the DATASET A while the bottom panels refer to DATASET B.

Figure 2 show the log-log plot of the cumulative distribution function and maximum likelihood power-law fit for the centrality measures employed in the analysis. In the figure we separate the cumulative distributions of such measures for the defaulted and non defaulted institutions. In all the measures and for both the two types of data set we observe different scalings of the power-law exponents for the institutions belonging to the defaulted set and for the sound ones, this suggests that, potentially, the centrality measures that account for nodes' importance are useful variables for discriminating between companies.

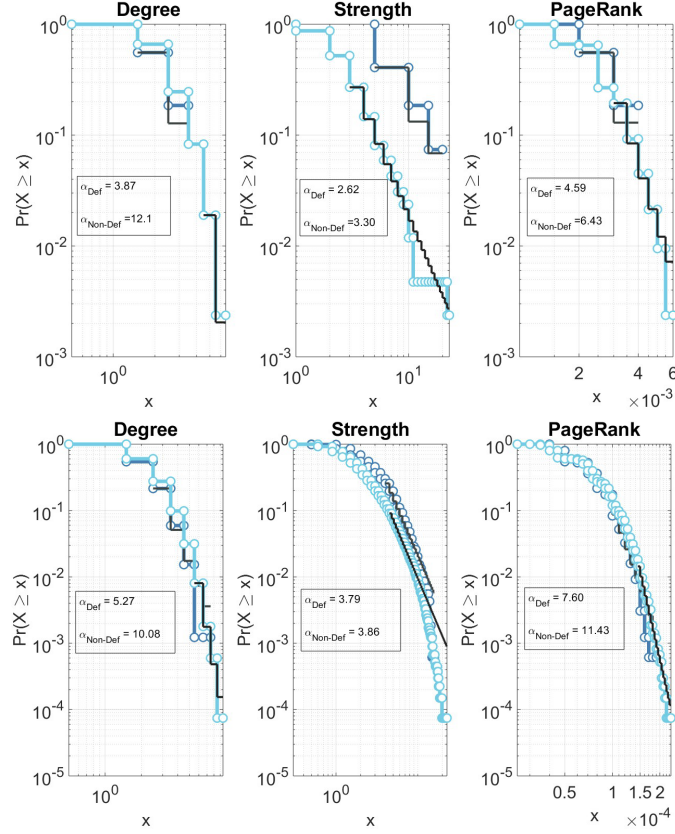


Figure 2: Centrality measure distributions. The panels represent the distribution of the centrality measures separated according to the defaulted indicator, together with the corresponding power-law coefficient estimate. In the left panels we represent the degree distributions, the central panels refer to the strength distributions while the right panels encompass the PageRank distributions. Moreover, the upper panels refer to the DATASET A while the bottom panels refer to DATASET B. The different values of the scaling coefficients related to the distribution of defaulted and non defaulted institution suggest their potential values for discriminating between companies.

Information concerning the community structure of the networks are used, together with centrality measures, to provide synthetic topological measurements at node level. Such variables are then embedded into credit scoring models to assess whether they contain relevant information useful for forecasting institutions default.

Performance improvements for all the tested models and for both data sets are reported in Table 1. The table summarizes the values of the measures employed to assess the predictive gain of the network augmented credit scoring models. We report, the area under the ROC curve (AUC), the area under the PR curve (AUPR), the model accuracy (ACC), the F1-score (F1) and Net Reclassification Improvement (NRI). From the results collected in Table 1, it is quite clear that the inclusion

DATA-SET A	AUC		AUPR		ACC		F1		NRI	
	Base Network		Base Network		Base Network		Base Network		Nri	P-val
Logistic Regression	0.7252	0.8021	0.1827	0.2653	0.9376	0.9510	0.9687	0.9688	0.5805	(0.09)
Discriminant Analysis	0.7197	0.7197	0.2590	0.2766	0.9443	0.9376	0.9684	0.9684	-0.0579	(0.54)
Naive Bayes	0.7404	0.7447	0.2358	0.2472	0.8730	0.8775	0.9655	0.9656	-0.039	(0.10)
Support VM	0.6014	0.7160	0.1361	0.1556	0.9398	0.9420	0.9689	0.9687	0.202	(0.55)
Decision Trees	0.7160	0.7178	0.2340	0.2416	0.9354	0.9376	0.9678	0.9679	-0.039	(0.20)
DATA-SET B	AUC		AUPR		ACC		F1		NRI	
	Base Network		Base Network		Base Network		Base Network		Nri	P-val
Logistic Regression	0.8155	0.8229	0.3434	0.3418	0.9034	0.9036	0.94208	0.94196	0.004	(0.79)
Discriminant Analysis	0.8011	0.8126	0.2942	0.3038	0.8888	0.8931	0.9400	0.9401	0.011	(0.08)
Naive Bayes	0.8064	0.8090	0.3097	0.3038	0.8029	0.8063	0.9310	0.9313	-0.015	(0.00)
Support VM	0.6997	0.7543	0.1615	0.2470	0.8932	0.8965	0.9424	0.9422	0.353	(0.00)
Decision Trees	0.7124	0.7097	0.1924	0.1899	0.8742	0.8705	0.9387	0.9384	-0.026	(0.00)

Table 1: Summary Statistics of non-parametric analysis. Summary statistics of the non-parametric analysis. From the left to the right: area under the ROC curve (AUC), area under the PR curve (AUPR), model accuracy (ACC), F1-score (F1) and Net Reclassification Improvement (NRI). For each measure and for all the tested models we reports the results obtained by the base-line scenario and for the network-augmented configurations. Moreover, for the NRI we report the p-value of the statistic. Notice how the inclusion of topological variable sensible increase the performance of the models, especially for the logistic regression and support vector machine classifiers.

of topological variables describing institutions centrality in the similarity networks and the community structure composing such networks increases the predictive performance of most of the credit scoring models even if the forecasting gain obtained differ from model to model, depending also from the data set employed. In particular, we observe an increase of the AUC and AUPR values especially for the logistic regression and for the support vector machine models while gain are more contained for the discriminant analysis, naive Bayes models and decision tree. Concerning models accuracy, ACC and F1- score measures are less sensitive to the inclusion of topological variables even if with some exception, as in the application of the logistic regression to the first data set.

Finally, since our main aim is to evaluate whether the inclusion of network measures could improve the predictive performance of some classes of credit scoring models we also report the Net Reclassification Improvement (NRI). This measure has been developed by Pencina et al. (2008) and explicitly accounts for the gain obtained by the inclusion of other dependent variables in scoring models. Results reveal increment of the NRI vales and therefore in the forecasting performance of the logistic regression and support vector machine models in the first data set and in the same two models and also for the discriminant analysis in the second data set. Despite the fact that non all of these increments are statistically significant, the additional information provided by topological measures improved the classification for a net of 58 per cent of firms with the default events, with no net loss for non-events in the case of logistic regression (DATA-SET A) and for a net of 35 per cent in the case of support vector machine (DATA-SET B).

References

1. Abellán, J. and Castellano, J. A comparative study on base classifiers in ensemble methods for credit scoring. *Exp. Sys. Appl.*, 73:1-10, (2017).
2. Blondel, V., Guillaume, J.L. Lambiotte, L., and Lefebvre, E. Fast unfolding of communities in large networks. *JSM*, 2008(10):P10008, (2008).
3. Bonacich, P. Some unique properties of eigenvector centrality. *Soc. Net.*, 29(4):555-564, (2007).
4. Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. *Com. Net. and ISDN systems*, 30(1-7):107-117, (1998).
5. Katz, L. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39-43, (1953).
6. Khandani, A., E., Kim, J. and Lo, A. Consumer credit-risk models via machine-learning algorithms. *JBF*, 34(11):2767-2787, (2010).
7. Khashman, A. Credit risk evaluation using neural networks: Emotional versus conventional models. *Appl. Soft Comp.*, 11(8):5477-5484, (2011).
8. Kleinberg, M. Authoritative sources in a hyperlinked environment. *JACM*, 46(5):604-632, (1999).
9. Lessmann, S., Baesens, B., Seow, H., and Thomas, L. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *E.J. Oper. Res.*, 247(1):124-136, (2015).
10. Pecora, N., Kaltwasser, P., R., and Spelta, A. Discovering sifis in interbank communities. *PloS one*, 11(12):e0167781, (2016).
11. Pencina, M., J., D'Agostino, R., and Vasan, R. Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Stat. Med.*, 27(2): 157-172, (2008).
12. Perra, N., and Fortunato, S. Spectral centrality measures in complex networks. *Phys. Rev. E*, 78(3):036107, (2008)
13. Spelta, A., Flori, A., and Pammolli, F., Investment communities: Behavioral attitudes and economic dynamics. *Soc. Net.*, 55:170-188, (2018).

Between hawks and doves: measuring central bank communication

Fra falchi e colombe: valutazione delle comunicazioni di Banca Centrale

Ellen Tobback, Stefano Nardelli, David Martens

Abstract We propose a Hawkish-Dovish (HD) indicator that measures the degree of 'hawkishness' or 'dovishness' of the media's perception of the ECB's tone at each press conference. We compare two methods to calculate the indicator: semantic orientation and Support Vector Machines text classification. We show that the latter method tends to provide more stable and accurate measurements of perception on a labelled test set. Furthermore, we demonstrate the potential use of this indicator with several applications: we perform a correlation analysis with a set of interest rates, use Latent Dirichlet Allocation to detect the dominant topics in the news articles, and estimate a set of Taylor rules. The findings provide decisive evidence in favour of using an advanced text mining classification model to measure the medias perception and the Taylor rule application confirms that communication plays a significant role in enhancing the accuracy when trying to estimate the bank's reaction function.

Abstract Nel documento viene presentato l'indicatore Hawkish-Dovish (HD) che misura il grado di durezza ("hawkishness") o morbidezza ("dovishness") nella percezione mediatica del tono della comunicazione alle conferenze stampa che seguono le decisioni di politica monetaria della BCE. Vengono confrontati due metodi per calcolare l'indicatore: il primo è basato sull'orientamento semantico mentre il secondo sulla classificazione del testo sulla base della tecnica basata sulle macchine a vettori di supporto (Support Vector Machine). L'evidenza prodotta nel documento indica che quest'ultimo metodo tende a fornire misurazioni più stabili e accurate della percezione nella predizione su un insieme campione. Inoltre, l'uso potenziale di questo indicatore viene mostrato in relazione a diverse applicazioni: un'analisi di correlazione con vari di tassi di interesse ufficiali e di mercato, un esercizio basato sull'Allocazione di Dirichelet latente (Latent Dirichlet Allocation) per rilevare gli argomenti dominanti negli articoli, e la significatività di tale indicatore in un insieme di regole di Taylor. I risultati forniscono prove decisive a favore dell'utilizzo di un modello avanzato di classificazione di analisi testuale ("text mining") per misurare la percezione dei media e l'applicazione della regola

di Taylor conferma che la comunicazione svolge un ruolo significativo nel migliorare l'accuratezza nella stima della funzione di reazione della banca centrale.

JEL codes: C02, C63, E52, E58

Key words: Monetary policy, communication, data mining, quantitative methods

1 Non-technical summary

In recent years communication became increasingly important in central banks. In particular, after the financial crisis, communication has increasingly qualified as a genuine policy tool able to steer interest rates in financial markets and drive expectations on the course of monetary policy. With official interest rates approaching zero and thereby reducing their effectiveness, various forms of forward guidance (i.e. a verbal commitment on the future course of monetary policy unconditional or conditional to some economic event) were added to the standard monetary policy toolkit existing out of interest rates and refinancing instruments for the banking sector. The growing relevance of communication in the conduct of monetary policy was mirrored by a rising interest of academicians and practitioners. In particular, a branch of economic research increasingly focused on the role of communication in adding valuable information besides what is already contained in macroeconomic variables, and in revealing policy makers preferences on the course of monetary policy with a view to enhance predictability. If the traditional approach consists in analysing how information events impact financial market developments and expectations of future policy moves, more recently analysis shifted towards analysing the language used by the central bank in its statements and how such message is perceived by its stakeholders. This paper contributes to the latter with a numerical indicator, called HD index (after the initials of hawkish and dovish), derived from media reports on the ECB press conference. Combining concepts and techniques developed in the context of computational linguistics and data mining, the indicator extracts relevant information on ECB monetary policy as reported by external observers and may therefore be interpreted as how media perceive the central banks monetary policy messages. For practical reasons, the perception is expressed on a numerical interval between -1 (most dovish) and +1 (most hawkish). In other words, the indicator indicates whether the perceived tone on monetary policy communication is predominantly on the tightening side (hawkish perception) or rather on the loosening side (dovish perception). Although the approach to quantify communication is not new, the approach proposed in this paper is original in various dimensions. First, the indicator does not measure directly official central bank communication but how such communication is received and interpreted with the crucial support of data mining techniques. Second, two different techniques are employed to compute the HD index: one based on the semantic orientation (SO) and

Between Hawks and Doves: measuring central bank communication

the second on text classification using Support Vector Machines (SVM). The former, the most commonly used by researchers, measures how often the ECB is mentioned in a news article together with a number of given hawkish and dovish words or expressions, while the latter, computationally more complex, uses a classification model to predict the tone of an article. Both methods are applied to a data set of around 9,000 articles published between January 1999 and March 2016 in order to assess which methodology produces better results. Based on various criteria (event analysis, correlations with actual interest rates and classification method) the SVM methodology tends to produce better and more reliable results. Third, in addition to its superiority on the SO, the SVM classification model can be used to analyse the terms most frequently employed by media in relation to a likely future course of monetary policy. Finally, an expanded Taylor rule framework including the HD index alongside traditional variables measuring inflation expectations and economic slack, is presented. Results suggest that the significance of the HD index as well as a relatively better fit confirm the positive role of ECB communication in enhancing the accuracy when trying to estimate the bank's reaction function, and thus that, on average, the ECB messages are correctly understood by its media watchers.

This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB

Ellen Tobback, Stefano Nardelli, David Martens.

Ellen Tobback University of Antwerp, Antwerp, Belgium;
email: ellen.tobback@uantwerpen.be

Stefano Nardelli European Central Bank, Frankfurt am Main, Germany;
email: stefano.nardelli@ecb.europa.eu

David Martens University of Antwerp, Antwerp, Belgium;
email: david.martens@uantwerpen.be

New methods and data sources for the population census

Nuovi metodi e fonti per il censimento della popolazione

Paolo Valente

Abstract For many years, all countries conducted their population census with the same traditional approach, based on the direct collection of information on individuals and households using census forms. However, a census conducted in this way is a very complex and expensive operation. Starting in the 1960s, some countries in Northern Europe developed register-based censuses using administrative data. Then in the 1990s, some countries developed alternative methods to conduct the census, combining administrative data and a partial data collection, with the objective to limit the costs, produce good quality results and, in some cases, increase the census frequency. This paper presents a review of the methods and data sources used by European countries to conduct their census.

Abstract Per molti anni, tutti i paesi hanno eseguito il censimento della popolazione utilizzando lo stesso approccio tradizionale, basato sulla raccolta diretta delle informazioni su individui e famiglie utilizzando questionari censuari. Tuttavia, un censimento condotto in questo modo è un'operazione molto complessa e dispendiosa. A partire dagli anni 1960, alcuni paesi in Nord Europa hanno sviluppato dei censimenti basati sui registri amministrativi. Negli anni 1990, alcuni paesi svilupparono metodi alternativi per condurre il censimento, combinando dati amministrativi e una raccolta parziale di dati, con l'obiettivo di contenere i costi, produrre risultati di buona qualità e, in alcuni casi, aumentare la frequenza dei censimenti. Questo articolo presenta una rassegna dei metodi e delle fonti utilizzate dai paesi europei per condurre il censimento.

Key words: Administrative sources, census, population, registers, surveys.

¹ Paolo Valente, United Nations Economic Commission for Europe (UNECE); email: paolo.valente@un.org

1. The traditional census approach

Population censuses have been conducted since the ancient world, when they were mainly for military and fiscal purposes. In the 18th century various countries started conducting complex censuses at regular intervals, in order to collect general information on the population, and not only for military or fiscal purposes. In the 19th century all European states were regularly conducting population censuses, normally every 10 years (Livi Bacci, 1981).

At that time, the population census was conducted in all countries adopting basically the same approach - often referred to as “traditional census” - based on the direct collection of information on individuals and households using paper questionnaires. With this approach, the national territory is divided into many census areas, each of them assigned to a census enumerator who is responsible for the delivery and collection of the census forms. In some countries the forms are completed by the enumerator based on information provided by the households, while in other countries the forms are completed directly by the household members.

1.1 Shortcomings of the traditional census

Conducting a traditional census with paper forms is a hugely expensive and complex operation. The main cost item is for the very large temporary work force (enumerators, supervisors, etc.) that has to be recruited and trained, and then has to work for a few weeks or longer periods. Moreover, a huge number of census forms must be printed, distributed, and collected. Then, the data have to be entered (manually or using scanners), processed, edited and finally tabulated and published.

Given the huge number of forms, the operations of collection, control and processing of the results are particularly long and complex. For this reason, and because of the high costs involved, the traditional census is generally conducted every 10 years. So, under the traditional approach both the frequency and timeliness of the census results are not satisfactory for many users.

In recent years, additional problems associated with the traditional census have emerged in many countries, including the increasing reluctance of the population to participate in the census and the difficulty of enumerating certain population groups, particularly those characterized by high mobility and multiple residences (e.g. students, young professionals) (Valente, 2015).

1.2 Use of long form and short form

Various solutions were developed by countries to deal with the problems associated with the traditional census. One of them, adopted for instance by Canada and the United States since the 1970s, consists of using two different forms: a long form used

to collect detailed information from a sample of the population, and a short form used for the majority of the population, to collect only very general information required for the population count.

ADVANTAGES: With this method, the amount of information collected and processed is substantially reduced, resulting in lower costs, reduced complexity of the census operations, and more timely publication of the results.

DISADVANTAGES: The information present in the long form is available only for a sample of the population. Therefore, for those variables the level of detail is limited both for small areas and for small population groups.

1.3 Internet response

Starting in the late 1990s, some countries developed the internet response option as part of a traditional census, with the objectives of maintaining or increasing coverage, improving data quality, shortening data processing time, and eventually reducing costs in the long term, if a significant proportion of the respondents use the internet.

In 2000-2001 a few countries offered the internet response option, in general to test this solution, with low take-up rates (the highest was 4% in Switzerland). In 2005-2006 the internet response was adopted more extensively in some countries, with higher take-up rates (18% in Canada). In 2010-2011 the internet response was adopted in many countries, particularly in Europe, with very high take-up rates in Estonia (67%) and Canada (54%) (Valente, 2012).

ADVANTAGES: Improved coverage among selected population groups; improved data quality thanks to the possibility to include logic controls in the electronic questionnaire; faster data processing; reduced costs only when a high proportion of the respondents use internet.

DISADVANTAGES: It requires certain levels of internet penetration and willingness by respondents to complete the census questionnaire online; additional development costs for electronic questionnaire and IT infrastructure; security risks associated with use of internet (e.g. hackers, risk of IT problems due to high volume of connections); need to track responses from multiple modes (see section 5).

1.4 Rolling census

An original solution to deal with the shortcomings of the traditional census was developed in France and implemented starting in 2004.

Under this approach, the census is conducted as a cumulative continuous (or “rolling”) survey over a long period of time rather than on a relatively short time period. In France a five-year cycle was adopted for the rolling census, and two different strategies are used: small municipalities (population under 10,000) are divided into five groups, and a full census is conducted each year in one of the groups;

in large municipalities (population 10,000 and above) a sample survey covering 8% of dwellings is conducted each year. At the end of the five-year cycle, about 70% of the country's population is enumerated. This is enough to guarantee robust information at the level of municipality and neighbourhoods, according to the French statistical institute INSEE. The census results are based on rolling averages calculated over the five-year cycle, and are updated yearly (Clanché, 2013).

ADVANTAGES: Data collection work and costs are distributed over time; census staff is constant over time, with no need to recruit and train new staff for each census; improved frequency of census results, as revised estimates can be produced every year.

DISADVANTAGES: The method is relatively complex to implement; corrections are needed to take into account migration movements between different parts of the countries (as a result of migration, some persons might be enumerated twice in the five-year period, and others might never be enumerated).

2. The register-based census²

A totally different approach from the traditional census was developed by the Nordic countries from the 1970s, replacing the census enumeration by the use of administrative data coming from various registers (population register, cadaster, social security, etc.) through a matching process, making use of a personal identification number. This approach, used for the first time in Denmark for the 1981 census, was progressively adopted by all other Nordic countries, and more recently by many other countries in Europe and beyond.

Some countries conduct a register-based census in which data are derived from existing sample surveys for selected variables. The first example was the so-called "Virtual census" conducted in the Netherlands in 2001 (Statistics Netherlands, 2004). Since the registers did not include all the necessary information for some of the economic characteristics, they were integrated with results from the labour force survey (LFS) in order to produce census data. In 2011, this approach was adopted by Belgium, Iceland and, again, in the Netherlands.

REQUIREMENTS: Registers must include all census variables, with sufficient levels of coverage and quality; the public must accept the register-based statistical system; the legal framework must allow the access, use and matching of register data, including personal identification numbers; there must be good cooperation between the statistical office and the authorities responsible for the registers; setting up and maintaining a statistical system based on registers normally follows a strategic decision by the statistical office (concerning not only the census but the overall statistical production) and requires important initial investments and a very long development time.

² More detailed information on register-based censuses is available in UNECE (2007, 2015, 2018)

ADVANTAGES: Once a good quality system of statistical registers has been set up, data for the census and for other statistical activities can be produced at a limited cost and with relatively limited work; there is no burden on respondents; data are potentially available every year.

DISADVANTAGES: Availability, coverage and quality of census data depends on data available in registers; register data is collected for administrative purposes, so definitions, classifications, etc. may not fully meet statistical requirements.

3. The combined census³

Many countries have population and other registers that potentially could be used for the census, but the coverage and/or data quality are not sufficient to use them as the only source to produce census data. It may be the case that some key census variables are not available in either the registers or existing surveys. Instead of continuing with the traditional census, for the 2000 round some of these countries used register data combined with a limited field data collection in order to produce the census results. Different approaches to this “combined census” exist, depending on the data sources used, and the type of field data collection. The two main approaches are presented in this section.

3.1 *Combining data from registers and an ad hoc sample survey*

Under this approach, data from registers are combined with data from a sample survey conducted ad hoc for the census. The survey has two objectives: 1) to evaluate the coverage and quality of the population, address, or other registers; and 2) to collect information on topics that may not be covered in registers, or for which the coverage and quality of registers is not sufficient. This method was adopted in 2008 by Israel, and in 2011 by other countries including Poland, Spain, Switzerland and Turkey.

The new “permanent census” launched in 2018 in Italy can be considered a mix of this approach and the rolling census, since administrative data are used in combination with data collected annually through specific sample surveys.

REQUIREMENTS: Same requirements as for register-based censuses.

ADVANTAGES: Efficient use of available data; possibility of assessing the quality of register data, and consequently being able to adjust population counts derived from it; reduced costs and complexity compared to a traditional census; limited burden on respondents (due to sampling).

DISADVANTAGES: For variables obtained from the sample survey, the level of detail is limited both for small areas and for small population groups, due to the sampling.

³ More detailed information on register-based censuses is available in UNECE (2015, 2018)

3.2 Combining data from registers and full enumeration

This approach is similar to the previous one, but data from registers are combined with data for selected variables obtained from a full enumeration (instead of an ad-hoc sample survey). For these variables detailed information will be available, but the cost and complexity of the data collection will be higher compared to the previous approach. Several countries in the European Union used this approach for the 2011 census, including Germany, Estonia, Latvia and Lithuania (UNECE, 2019a).

REQUIREMENTS: Same requirements as for register-based censuses.

ADVANTAGE (compared to the previous approach): More detailed information on the variables not adequately covered in the registers.

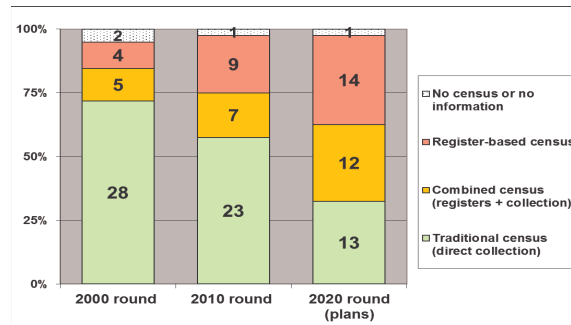
DISADVANTAGES: More expensive and complex than other types of register-based censuses because of the full field enumeration, but still less expensive than a traditional census because of efficiencies in field operations made possible using register data; heavy response burden on the public, due to the full enumeration.

4. Evolution of census methods over time in Europe

As described in the previous sections, over the past decades in Europe an increasing number of countries have adopted non-traditional census methods. Figure 1 presents the number of European countries that used the different census methodologies in the 2000 and 2010 census rounds, and the plans for the 2020 round⁴.

The figure shows clearly that the proportion of countries adopting an alternative method (register-based or combined census) increased significantly from about one quarter in the 2000 round to two thirds in the 2020 round (based on the plans communicated by the countries).

Figure 1: Number of European countries by census method adopted – 2000, 2010, and 2020 rounds⁵



Source: Information available at the United Nations Economic Commission for Europe⁴

⁴ Source for the 2020 round plans: UNECE (2019b). Online table regularly updated by UNECE.

⁵ According to the United Nations World Programmes on Population and Housing Censuses, the 2000, 2010 and 2020 census rounds refer to the periods 1995-2004, 2005-2014 and 2015-2024 respectively.

Only five of the countries that plan to conduct a traditional census in the 2020 round are EU member countries. In fact, the large majority of EU and EFTA countries have adopted or are considering the adoption of an alternative census method based on administrative sources. This reflects the EU “Strategy for the post-2021 census” which should include annual population data collections based on administrative data sources (starting in the mid-2020s) and larger multi-annual and/or decennial data collections in the form of the recurring decennial census (from 2031) (Eurostat, 2018).

5. Multi-mode approaches: challenges and future developments

Many of the approaches described above imply the use of data obtained from multiple sources and/or with multiple collection modes. In a traditional census, for instance, countries may combine internet response, CAPI in some parts of a country, and paper-based interviews in other parts, typically in remote regions. In a combined census, data can be derived from administrative sources and an ad hoc field collection.

Multi-mode approaches have the advantage of making use of existing data sources and of technologies, even when they are available only in part of a country (e.g. using tablets only in urban areas). This may increase the census efficiency in terms of improved response rate, coverage or data quality, reduced time needed for data collection and processing, and possibly reduced costs.

However, adopting a multi-mode approach also implies a number of challenges and risks. It is necessary to set up a response tracking system able to record in real time the responses received from the different modes, and share it with field workers. Otherwise, for example, an enumerator could go and visit a household that has already responded by internet, with unnecessary costs and potential risks of double counting.

Whatever census method is selected, it is important that planning starts very early and that all processes – particularly those involving innovations – are tested carefully before the final decision is taken about the methodology/technology to be adopted.

A careful risk management analysis should be conducted – particularly when a technology is adopted for the first time – and if possible, contingency plans should be prepared in advance, to be able to react in case of unexpected problems.

Finally, countries should develop methods to assess the quality of the different data sources, to be able to provide the users with an evaluation of census data quality.

For the future, it can be expected that the trend of moving away from the traditional census and adopting alternative methods based on administrative data will continue. However, some countries – namely those where administrative sources are not developed or not suitable for statistical use – may continue with the traditional census.

A possible development for future censuses is the use – in combination with other sources – of “big data”. Various countries have been testing the use of big data with a view to using such sources for the 2021 census. Spain, for instance, tested the use of data from mobile phone companies to support the production of census data on commuting and mobility, by identifying the enumeration areas where respondents live, and those where they work or study. Spain also tested the use of data on

electricity consumption to distinguish dwellings that have been occupied for all or at least part of the year from those that are vacant (Vega et al., 2016).

Although the use of big data for censuses (and for official statistics generally) may be promising in the long term, there are still many methodological, legal, quality and IT challenges. Subject to positive results of the current work in this field, big data could be used by some countries to support censuses after the 2020 round (CSO of Poland, 2016).

References

1. Clanché, F.: The French rolling census, ten years after its launch. UNECE Group of Experts on Population and Housing Censuses, Geneva, 30 September – 3 October 2013 (2013). https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2013/census_meeting/24_E.pdf Cited 20 Feb 2019
2. CSO of Poland: Is it possible to use Big Data in the 2020 Census Round?, UNECE Group of Experts on Population and Housing Censuses, Geneva, 28 - 30 September 2016 (2016), http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2016/mtg1/6_Spain_ENG.pdf Cited 20 Feb 2019
3. Eurostat: Plans for the 2021 EU census programme and for the development of European population statistics beyond 2021: an update. UNECE Group of Experts on Population and Housing Censuses, Geneva, 26–28 September 2018 (2018). http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2018/Meeting-Geneva-Sept/ECE.CES.GE.41.2018.23_Item_4_YT.pdf Cited 20 Feb 2019
4. Livi Bacci, M.: Introduzione alla demografia. Loescher editore, Torino (1981)
5. Statistics Netherlands: The Dutch Virtual Census of 2001 - Analysis and Methodology. Statistics Netherlands, Voorburg/Heerlen (2004). <https://www.cbs.nl/-/media/imported/documents/2005/17/b-57-2001.pdf?la=en-gb> Cited 20 Feb 2019
6. UNECE: Register-based statistics in the Nordic countries - Review of best practices with focus on population and social statistics, United Nations, New York and Geneva (2007). http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf Cited 20 Feb 2019
7. UNECE: Conference of European Statisticians Recommendations for the 2020 Censuses of Population and Housing, United Nations, New York and Geneva (2015). http://www.unece.org/fileadmin/DAM/stats/publications/2015/ECECES41_EN.pdf
8. UNECE: Guidelines on the use of registers and administrative data for population and housing censuses, United Nations, New York and Geneva (2018). <http://www.unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20184.pdf>
9. UNECE: Wiki table on the 2010 census round (2019a): <https://statswiki.unece.org/display/censuses/2010+Population+Census+Round> Cited 20 Feb 2019
10. UNECE: Wiki table on the 2020 census round (2019b): <https://statswiki.unece.org/display/censuses/2020+Population+Census+Round> Cited 20 Feb 2019
11. Valente, P.: Use of the internet response for censuses of the 2010 round in the UNECE region. United Nations International Seminar on Population and Housing Censuses: Beyond the 2010 Round, 27-29 November 2012, Seoul, Republic of Korea (2012). <https://unstats.un.org/unsd/demographic/meetings/Conferences/Korea/2012/docs/s07-2-1-UNECE.pdf> Cited 20 Feb 2019
12. Valente, P.: Censuses: Current Approaches and Methods. In: James D. Wright (editor-in-chief), International Encyclopedia of the Social & Behavioral Sciences, 2nd edition, Vol 3, pp. 296–301. Elsevier, Oxford (2015)
13. Vega, J. L. Argueso, A., Teijeiro, C.: Three examples of innovative data sources in 2021 Spanish Census. UNECE Group of Experts on Population Censuses, Geneva, 28 - 30 September 2016 (2016), http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2016/mtg1/6_Spain_ENG.pdf Cited 20 Feb 2019

FinTech and the Search for "Smart" Regulation

Fintech e la ricerca di una regolamentazione "smart"

Silvia Vanoni

Abstract This paper intends to give an overview of the interlinkage between finance and innovation, of the ways in which digital technologies are changing financial activities on a global scale, the opportunities and risks that they offer and the approaches of legislatures vis-à-vis these new phenomena. Emphasis will be placed on the actions taken by the competent authorities of the European Union and of Italy.

Abstract *Il saggio intende offrire una panoramica dell'intreccio tra finanza e innovazione, del modo in cui le tecnologie digitali stanno modificando le attività finanziarie su scala globale, dei rischi e opportunità che offrono e dell'atteggiamento dei legislatori al cospetto di simile fenomeno. Speciale attenzione è dedicata alle iniziative dalle autorità comunitarie ed italiane.*

Key words: algorithm, automation, crowdfunding, crypto-assets, finance, high-frequency, innovation, legislation, regulation, securities, technology, trading.

1 Introduction

FinTech is the short term for financial technology, which refers to “technology-enabled innovation in financial services” [8].

In Italy, the industry of financial services is currently strictly and thoroughly regulated. The legislation in force stems from directives and regulations passed by the authorities of the European Union (EU), so that domestic laws are highly harmonized with those of the other EU member states. The process of harmonization has also been fueled by the creation of centralized supervisory authorities, such as

¹Silvia Vanoni, Full Professor of Business Law, Università Cattolica del Sacro Cuore, Milan, silvia.vanoni@unicatt.it.

the European Securities and Market Authority–ESMA, in the wake of the global financial crisis of 2008. The other main reference jurisdiction in the field of securities trading and financial services regulation is the comprehensive body of federal laws of the United States, initially passed in the early 1930s, as a reaction to the Wall Street Crash of 1929: such regulation, as subsequently amended and supplemented, is still the paradigm of the legal treatment of financial markets and services.

Traditionally, technology innovation and financial activities have been tightly interlinked and both worlds have mutually benefited from and fostered their respective developments. As a matter of fact, technology innovation on the one hand has enabled financial activities to extend their operations to a wider array of potential clients, leading to the progressive globalization of the industry and, on the other hand, it has sped up the pace of financial transactions and processes. Conversely, the development of technologies has been spurred by their extensive use by financial players¹.

In the past ten years, the fast progress of information technology has caused a massive irruption of innovation in the financial world, which has been fueled by the combination of two factors triggered by the global financial crisis of 2008: the need for cheaper ways to convey funds from investors to businesses; the raise of the level of mistrust felt by investors and businesses alike vis-à-vis traditional financial institutions (banks, broker/dealers and other financial intermediaries, insurance companies, the same authorities in charge of markets supervision).

Among the factors which are changing the features of the financial industry, one can mention, according to no specific listing criteria: crowdfunding, i.e. the use of web platforms, which replace traditional intermediaries to raise money from investors who wish to fund (as lenders and/or equity-holders) business activities²; the spreading of “crypto-assets”, such as virtual currencies, the most famous – or notorious- of which are the Bitcoins, which are based on blockchain and other forms of distributed ledger technology³; algorithmic trading and high-

¹ Among the many milestones that mark the so-far happy “marriage” between the financial services industry and technology innovation, one can single out the introduction of the telegraph and the laying of the transatlantic cable in the 19th century, or the introduction of Automated Teller Machines (ATMs) and the replacement of “physical” financial markets with digital ones in the 20th century [1,11].

² Since 2012 the phenomenon of crowdfunding for businesses has been subject to legislative actions aimed at governing its functioning and use, with the main purpose of protecting investors without stifling the development of this alternative source of fundraising. Following the lead of the US Jumpstart Our Business Startups Act of 2012, Italy was the first EU country to pass some legislation on equity-based crowdfunding (Legislative Decree no. 179 of 2012 and CONSOB Regulation no. 18592 of 2013). In the next few years, most of the other EU member states took action to provide some sort of bespoke regulation operating within their national boundaries [7]. In March 2019, the European Parliament passed a regulation on cross-border equity-based and lending-based crowdfunding for businesses (Regulation on European Crowdfunding Service Providers (ECSP) for Business). Both the national and the European sets of rules focus on two main points: the information to be made available to investors; the amount of investments to be allowed to be raised through each placement offer and the overall value of the investments that each person should be permitted to make through a platform: in both cases, some sort of maximum thresholds were established.

³ Tokens (or coins) and the other “crypto-assets”, like the thousands of virtual currencies in circulation, raise the fundamental problem of legal classification. They are digitally issued in return for some

frequency trading (i.e., a sub-type of algorithmic trading) in financial markets; the practice of robo-advising.

This paper intends to offer an overview of the challenges which the development of new technologies in the financial area is presenting to legislatures, by focusing mainly on the situation in the European Union and in Italy. To this end, the author will proceed as follows: section 2 will deal with the general legal strategies adopted by the competent authorities in relation to innovation in the financial field; section 3 will deal with two types of FinTech phenomena in which the human factor has been fully replaced with the functioning of predictive algorithms; section 4 will set out some final remarks.

2 The approaches of the legislatures.

The traditional stance of legislatures towards technological novelties is to remain neutral on whether or not they should be adopted and to focus instead on the outcomes of their use. In this connection, legislatures must assess if such new “factors” are exercising any significant impacts on the pre-existing balance among the interests held by the various stakeholders involved in any given kind of financial relationship (mainly: retail investors, financial intermediaries, financial markets, financial supervisory authorities) and on the overall legal framework to which such relationships belong. In the affirmative case, legislatures must re-adjust the balance, by implementing the necessary new rules. In other words, it is the duty of legislatures to detect and weigh potential opportunities and risks stemming from technology innovation and to decide if, how and when to take action, in order to fill in any regulatory gaps [12]¹.

In 2018, the European Commission, after acknowledging that “The rapid advance of Fintech is driving structural changes to the financial sectors”, set three main goals for its legal action, namely; “to harness rapid advances in technology for the benefit of the EU economy, citizens and industry, to foster a more competitive and innovative European financial sector, and to ensure the integrity of the EU financial system” [8, 4]. At the same time, the Commission decided on refraining from hasty legislation and adopted a very cautious approach, which consists of

consideration, they are usually placed through public offers called “initial coin offerings” (ICOs) and are traded in unregulated digital markets. The technology used to create and transfer them shows potential for being fruitfully applied also to the infrastructures of financial regulated markets. However, the risks that they entail are identified in “strong volatility (...) fraud and operational weaknesses and vulnerabilities at crypto assets exchanges” [8], which can lead to, inter alia, money laundering, terrorists financing, market manipulation and substantial losses for investors [6, 9]. For these reasons, in 2017 China decided to ban virtual currencies [12]. On the other hand, in January 2019 ESMA proposed to distinguish between crypto-assets which fall within the legal definition of financial instruments or any other regulated legal phenomena and those which do not. Regarding the former, some update of existing legislation may be appropriate; regarding the latter, the design of a bespoke regime on ICOs and tokens circulation is deemed to be necessary [6]. To this end, it is held that crowdfunding platforms can be suitable means to carry out such transactions [3].

¹ The timing factor is not irrelevant or secondary, since too early regulation of a new phenomenon could result in a waste of time and resources and stifle the development and spread of innovation to a level high enough to produce assessable effects [8].

introducing some legal updates to address the most immediate risks and of closely and constantly monitoring the developments in the sector for all the other aspects.

Currently, all legislatures in the world are weighing the impact of FinTech on their respective legal environments.

Some peculiarities of the new wave of innovation are undisputed: (i) its pace is faster; (ii) the main actors often tend to be different from those of the past: instead of traditional institutions (banks, financial intermediaries, financial markets, which are already subject to specific set of rules governing their organizations and activities), many new players are now rooted in the IT industry; as a matter of fact, some of the mentioned novelties have the purpose to set aside the traditional actors and replace them with technology firms (see for instance the use of crowdfunding platforms or the spreading of virtual currencies); (iii) from a legal viewpoint, it is very difficult to classify some of the products of FinTech, namely most of the crypto-assets which are issued and circulating, and consequently to verify whether such assets fall within the scope of any pre-existing regulation or if they require the passing of new legislation; (iv) extensive digitalization aggravates the general problem of data management and protection [12].

3. A look at some FinTech phenomena and at related legislative reactions: high-frequency trading (HFT) and robo-advice.

HFT and robo-advice relate to different financial services but both entail the substitution of human activities for the operations of predictive algorithms. Whilst by now HFT has become a common practice of trading venues, robo-advice has started to be adopted much more recently.

3.1 High-frequency trading.

HFT refers to trading activities in financial markets by intermediaries. It is not a new trading strategy, but a technologically advanced way to implement traditional strategies [2, 14]. It has become an increasingly common practice in the US and later on in the EU financial markets for the past 20 years¹.

HFT is a type of algorithmic trading, which occurs when decisions about securities trading are made and performed by algorithms in milliseconds as designed reactions to pre-determined types of information: publicly available information, fundamental value information (i.e., about the securities traded and their issuing entities) and, mostly, information about third parties' trading strategies which can be detected from the analyses of the orders entered by the other traders and anticipated through HFT².

¹ Nowadays, HFT is estimated to account for over 50% of trading volumes in equity markets in the US and for up to 40% of volumes in the markets of EU member states [2,14].

² HFT is defined in detail by article 4(1)(40), EU Directive 2014/65/EU (MiFID II) and by the ensuing Italian legislation (article 1, para. 6-septies, Legislative Decree no. 58 of 1998, a.k.a. Consolidated Code on Finance - CCF).

The main opportunities associated with HFT are that it is faster than the other traders –human or automated– in taking advantage of new information; it decreases operational costs; it limits holding risks; it injects liquidity into markets, although for a short length of time, since positions are usually closed by the end of each day by executing or cancelling the orders made; it reduces the benefits which can be extracted by inside traders, through an early detection of and fast reaction to their strategies.

Conversely, the main risks associated with such practice are that it is pro-cyclical and can lead to “flash crashes”; it can cause market manipulation, because of the speed and amount of orders which can be entered and cancelled each day; it increases market volatility; it may result in overreactions to inaccurate/misleading information, moving the market towards “wrong” directions [2, 14].

In the past few years, the European and Italian legislatures have provided for the application of ad-hoc strict regulation to this type of trading, aiming at restraining its use and strengthening market stability and integrity.

For instance, investment firms which carry out HFT must keep accurate and time sequenced records of all orders entered and cancelled; market authorities have access to information about the algorithms which are used and the ways in which they are tested and monitored by firms; the same authorities must also monitor the impact of HFT on market stability; trading venues are required to be equipped to reject orders that exceed certain thresholds of price and volume and can be authorized to apply higher commissions to orders which are cancelled by the end of the day (article 17, EU Directive 2014/65/EU (MiFID II); articles 65-sexies and 67-ter, Italian Legislative Decree no. 58 of 1998, a.k.a. the Consolidated Code on Finance - CCF). Some scholars also point out that additional measures should be adopted in order to further “level the playing field” between HFT and human traders in relation to the access and use of relevant information [2].

3.2 Robo-advice.

Advice on financial instruments is one of the areas in which automation has been gradually expanding its reach in recent years, although it is far less common than HFT. Automated advisers, also referred to in short as “robo-advisers” “use innovative technologies to provide discretionary asset managerial services to their clients through on-line algorithmic-based programs” [13].

This type of financial services are provided through interactive web platforms to which clients can connect. Robo-advising services to investors can be classified in two different groups: fully automated advice, in which every step of the advisory process is performed through a platform, with no interaction with any human element; partially automated or “hybrid” advice, in which some level of human interaction is still contemplated¹.

¹ A third type of advisory activity is the so called “robo4advice”, which consists of the use of software programs by human financial experts to assist them in providing advisory services to their clients. In this situation, there is no direct impact on the structure of the relationship between the advisers and their clients; the pattern of robo4advice is a type of B2B service which does not qualify as a regulated financial service under applicable rules and for this reason it will not be considered in this analysis.

Under Italian law, investment advisory activities are subject to specific rules of the Consolidated Code on Finance, implementing MiFID II, whenever they fall within the scope of the definition set out by article 1, para. 5-septies CCF, that is: “the provision of recommendations tailored to a client following a request or on the initiative of the provider of the service, concerning one or more transactions relative to financial instruments”; in other words, bespoke recommendations of one or more transactions on securities. Professional advice having such features can be provided only by financial intermediaries or other entities who must be authorized by CONSOB and operate in compliance with strict rules of conduct aimed at investors’ protection.

The main principle governing the relationship client-adviser is that the recommended investment must be suitable to the client, in relation to the amount and quality of her wealth and her specific needs and expectations at the time when the investment is decided on. In order to comply with this principle, advisers must first of all obtain from their clients a fairly thorough set of information. To this end, at the inception of the relationship each client is required to fill in a questionnaire, which includes a test aimed at assessing the level of financial education of the investor to which the recommendation must be addressed. Such set of information should be kept updated in order to make sure that recommendations always meet the suitability requirement. Given that the existing rules have been designed by having in mind services performed by human beings, when the advisory activity is effected through an automated process, different opportunities and risks may emerge.

The opportunities offered by this new way of providing investment advice are first of all the reduction of the costs of the service, by making it accessible to a wider range of investors; secondly, the lack of or limited human interaction is likely to eliminate or at least reduce the risks connected to conflict of interest situations; finally, the replacement of human judgement and experience with the operations of some algorithm eliminates the chance of human error.

Specific risks are equally detectable in the various steps of the advisory process. Some of them pertain to the phase of information collection about the prospective investors through the questionnaire: for instance, an investor might not understand properly some questions, or might give incomplete answers, or might be unable to disclose crucial pieces of information owing to the structure of the questionnaire. Additional risks can be detected in the stage of formulating a recommendation and they can be linked to the inappropriate design of the algorithms or the malfunctioning of the software [10].

These issues have been analyzed by ESMA in guidelines addressed to national competent authorities and firms [5]. Said Authority emphasizes, *inter alia*, the importance of the information provided to clients about the use of automation (guideline no. 4); it recommends special care in designing the questionnaire aimed at gathering information from the clients and stresses the benefits of giving clients access to some type of human interaction; it also requires firms to ensure a constant

FinTech and the Search for “Smart” Regulation

testing, monitoring and updating of all the algorithms that they employ (guidelines nos. 82-100)¹.

In any case, it remains undisputed that the use of automated devices in financial advice services does not entail any limitation to the liability of advisers towards their clients for the performance of their obligations (e.g. article 167, para. 1, CONSOB Regulation no. 20307 of 2018).

4. Conclusive remarks on the systemic relevance of FinTech.

FinTech innovation is intertwined with and impacts on various aspects of financial activities, such as tradable products, services provided by intermediaries, market infrastructures: for this reason, its introduction has already caused and will further require some updating of the existing legislation.

However, some legal scholars suggest that one should not just focus on the issues raised by the adoption of each new type of technology, but look at the broader picture composed by the full set of FinTech innovation which has emerged in the past ten years and at the ways in which it has been introduced. This should lead to realize that we are not dealing with a series of individual, isolated changes, but with a complex structural phenomenon, which might affect the underpinning principles of the whole financial system. Consequently, a thorough re-design of the legislation of the financial sector should be taken into consideration, in order to be able to exploit the many opportunities offered by FinTech in a safe and “smart” way.

If policymakers failed to effectively adopt this comprehensive approach, the widespread use of the term “disruption” referring to the impact of FinTech innovation on the financial industry and its regulation might not be a mere hyperbole [11].

References

1. Arner, D., Barberis, J., Buckley, R.: The Evolution of Fintech: A New Post-Crisis Paradigm? (2015, last revised in 2016), available at <https://ssrn.com/abstract=2676553>.
2. Balp, G., Strampelli, G.: Preserving Capital Markets Efficiency in the High-Frequency Trading Era (2018), available at <https://www.ssrn.com/abstract=3097723>.
3. CONSOB: Le offerte iniziali e gli scambi di crypto-attività. Documento per la discussione (19 March 2019), available at <https://www.consob.it>.
4. ESMA: Response to the Commission Consultation Paper on Fintech: A more competitive and innovative financial sector, ESMA50-158-457 (7 June 2017), available at <https://www.esma.europa.eu>.
5. ESMA: Final Report. Guidelines on certain aspects of the MiFID II suitability requirements, ESMA35-43-869 (28 May 2018), available at <https://www.esma.europa.eu>.
6. ESMA: Advice. Initial Coin Offerings and Crypto Assets, ESMA 50-157-1391 (9 January 2019), available at <https://www.esma.europa.eu>.

¹ On the other hand, ESMA does not require advisers to provide their clients with any information about the features of such algorithms, which are protected by intellectual property rights, whilst the SEC suggests that some information about their structures be furnished to investors [13].

7. European Commission: Commission staff working document. Crowdfunding within EU Capital Market Union (3 May 2016), available at <https://www.ec.europa.eu>.
8. European Commission: FinTech Action plan: for a more competitive and innovative European financial sector (8 March 2018), available at <https://www.ec.europa.eu>.
9. Henderson, T., Raskin, M.: A Regulatory Classification of Digital Assets: Towards an Operational Howey Test for Cryptocurrencies, ICOs, and Other Digital Assets (2018, last revised 2019), available at <https://www.ssrn.com/abstract=3265295>.
10. Linciano, N., Soccorso, P., Lener, R. (editors): La digitalizzazione della consulenza in materia di investimenti finanziari, CONSOB Quaderni FinTech, no. 3 (January 2019), available at <https://www.consob.it>.
11. Omarova, S.T.: New Tech v. New Deal: Fintech As A Systemic Phenomenon (2018, last revised 2019), available at <https://www.ssrn.com/abstract=3224393>.
12. Schena, C., Tanda, A., Arlotta, C., Potenza, G.: Lo sviluppo del FinTech. Opportunità e rischi per l'industria finanziaria nell'era digitale, CONSOB Quaderni FinTech, no. 1 (March 2018), available at <https://www.consob.it>.
13. SEC: Investment Management Guidance Update. Robo-advisers, no. 2017-02 (February 2017), available at <https://www.sec.gov>.
14. Woodward, M.: The Need for Speed: Regulatory Approaches to High Frequency Trading in the United States and the European Union (2017, last revised 2018), available at <https://www.ssrn.com/abstract=3203691>.

An anisotropic model for global climate data

Un modello anisotropico per i dati climatici globali

Nil Venet and Alessandro Fassò

Abstract We present a new, elementary way to obtain axially symmetric Gaussian processes on the sphere, in order to accommodate for the directional anisotropy of global climate data in geostatistical analysis.

Abstract *Presentiamo un nuovo modo elementare per ottenere processi gaussiani assialmente simmetrici sulla sfera, al fine di accogliere l'anisotropia direzionale dei dati climatici nelle analisi geostatistiche su scala globale.*

Key words: climate data, geostatistics, Gaussian processes, anisotropy, sphere

1 Introduction

Following the increasing amount of measurements and computing power available, geostatistics has evolved in the last decades to consider massive and global scale datasets. This adaptation presents new challenges, such as taking into account the spherical nature of Earth, and the considerable non-stationarity of global datasets, while maintaining computing feasibility.

We focus in this article on Gaussian process based geostatistics for climate data. Isotropic model on the sphere, which assume that the data has the same statistical behavior in every direction, are well understood and the literature gives rich classes of covariances (see for example [4] and the reviews [9] and [13]).

However, climate data exhibits various nonstationary effects at the global scale. In particular, it is clear that the variability of climate variables is higher in the north-south direction, than in the east-west. Jones [10] notices this issue and defines axially symmetric Gaussian processes on the sphere, which are stationary in longitude only. He also characterizes their decomposition on the spherical harmonic basis, which Stein [15] truncates to a finite order to carry on a statistical analysis of a Total Ozone dataset at a global scale. While very flexible, Stein's model involves the estimation of a large number of coefficients, and still doesn't fit the local be-

haviour of the Ozone dataset. This calls for explicit axially symmetric covariances that relies on a small number of parameters. Such covariance functions are obtained through partial differentiation of isotropic processes by Jun and Stein [11, 12], and further studied by Hitczenko and Stein [6]. Recently Porcu et al. [14] proposed to modify variograms of isotropic covariances to obtain axially symmetric analogues, while Huang et al. [8] consider products of separated covariances on latitudes and longitudes.

Our approach consists in considering the product of an isotropic covariance with an additional covariance on latitudes. Compared to Huang et al.,[8], our method has the advantage of giving Gaussian processes that are continuous over the whole sphere, including the poles.

In Section 2, we give a description of two climatic datasets, which motivate our work and allow performance evaluation of models. Section 3 recalls some generalities on Gaussian processes on the sphere and introduce our model. In Section 4, we describe our ongoing work to compare model performances and properties to the existing literature.

2 Climatic datasets

In this section, we describe two important sources of climatic data, which motivates our study.

The radiosonde observations database from NOAA¹ ESRL² gathers worldwide radiosonde observations, that are considered as an important atmospheric observation standard. We propose to use it to assess model performances on real data.

We also consider the reanalysis ERA-Interim archive from ECMWF³, that does not present the drawbacks of real data (missing data, inhomogeneous spatial covering) but rather constitutes a resource whose coherence and completeness allows to extract tailor-fitted datasets to assess a virtually unlimited variety of hypothesis.

2.1 Radiosonde observations

As mentioned above, the NOAA ESRL database gathers worldwide radiosonde observations. There are around 900 ground stations performing radiosonde observations in the world. Ascending balloons equipped with measurement devices are launched, radio-transmitting the observations of meteorological variables about every few seconds to the ground.

¹ US National Oceanic and Atmospheric Administration

² Earth System Research Laboratory

³ European Centre for Medium-Range Weather Forecasts

These measurements are stored at a restricted number of altitudes levels, corresponding to pressure levels: there are 22 *mandatory levels* (from the ground up to 1 hPa) common to every sounding, to which are added a varying number of *significant levels* (28 on the average), that depend on the launch and correspond to important variations in measured temperature or dew point depression. Missing data at the highest mandatory levels are frequent as the balloon explodes between 30 and 35km of altitude.

There are typically two launches per day in a station, occurring at times close to 00 and 12 UTC, but additional and missing launches occur. Data is available publicly and can be downloaded from the NOAA ESRL servers.

2.2 The ERA-Interim reanalysis

Produced by the ECMWF, the ERA-Interim archive is a reanalysis of global atmospheric data from 1979, updated in real-time to nowadays.

This model output is built from the input of numerous meteorological datasets of various nature (mostly satellites measurements, but also radiosondes, land, boat, planes,... measurements), which are assimilated and extrapolated following a 4-dimensional variational analysis.

Era-Interim provides values for a considerable number of meteorological variables. Atmospheric variables are given at a vertical resolution of sixty pressure levels (from the ground to 0.1 hPa) on the horizontal reduced Gaussian grid N128, which is regular in latitude but has a variable longitudinal resolution to maintain a ground resolution of approximately 79km (see details in [7]). Time resolution is 6h. Surface variables are given with the same horizontal display and a time resolution of 3 hours. In both cases it is also possible to download data which have been extrapolated on regular latitude/longitude grids at various resolutions.

Advantages of this archive consist in its completeness, its coherence in the methodology and in the physical consistency of the data. From this, it is possible to tailor-fit datasets to a virtually unlimited number of applications.

Specifications of the archive are given in [1], while the reanalysis process is described in [3]. Data is available publicly and can be downloaded from the ECMWF servers.

3 Spatial Modeling

We briefly recall generalities about Gaussian process based geostatistics (for which we refer to Cressie [2] for more details) and Gaussian processes on the sphere before to introduce our new model.

3.1 Gaussian process geostatistics

The classical approach in geostatistics, given observed data $y_1, \dots, y_n \in \mathbb{R}$ at locations x_1, \dots, x_n in some space S consists in considering a Gaussian process $(Y_x)_{x \in S}$ indexed by S , and estimating the value of the variable of interest at a new location x^* by the Kriging estimator

$$\hat{Y}(x^*) = \mathbb{E}(Y(x^*) | Y(x_1) = y_1, \dots, Y(x_n) = y_n). \quad (1)$$

The quality of this prediction critically relies on the choice of the Gaussian process Y , whose statistical behavior should fit as much as possible the data of interest. The classical approach consists in selecting a Gaussian process from a parametric class Y_θ , where θ is a vector of parameters, which is typically done by maximizing the likelihood. The key asset of this approach is that prediction comes with a statistical model of the data, which allows in particular the computation of the conditional variance

$$\text{Var}(Y(x^*) | Y(x_1) = y_1, \dots, Y(x_n) = y_n), \quad (2)$$

that quantifies the prediction uncertainty at the location x^* .

Due to the $O(n^3)$ numerical complexity of both Kriging prediction and likelihood evaluation, this approach proves itself computationally heavy – if not unfeasible – on massive datasets. There exists an important literature on adaptations to large datasets. See for example [5] for a comparison of methods a single case study.

However in the majority of methods, the primary ingredient remain Gaussian processes, and finding classes of Gaussian processes that suit practical applications constitute an entire field of research.

Let us recall that the statistical properties of a Gaussian process are entirely characterized by its expectation and covariance functions. As the choice of a mean function is free from any constraints, in this article all Gaussian fields will be assumed centered. In this setting, considering a Gaussian process is equivalent to considering a covariance function.

In contrast with the mean function, not every function is admissible as a covariance. Let us recall that a function $F : X \times X \rightarrow \mathbb{R}$ is the covariance of a Gaussian process $(X_x)_{x \in X}$ if and only if it is symmetric and *positive definite*, that is

$$\forall \lambda_1, \dots, \lambda_n \in \mathbb{R}, \forall x_1, \dots, x_n \in X, \sum_{i,j=1}^n \lambda_i \lambda_j F(x_i, x_j) \geq 0. \quad (3)$$

In this case we will often say that F is a *valid covariance*.

The finding of classes of covariance functions that give Gaussian properties which fit data from application is the first, necessary step of Gaussian process modeling.

3.2 Gaussian processes on the sphere

In order to model variables on the Earth, we consider Gaussian processes indexed by the sphere \mathbb{S}^2 , which we will take of radius 1. To a given point x on the sphere are associated its longitude $\theta_x \in [-\pi, \pi]$ and latitude $\varphi_x \in [-\pi/2, \pi/2]$, in radians. Observe that θ_x and φ_x are well defined for every x distinct from the north or south poles, for which φ_x is $\pi/2$ (resp. $-\pi/2$) but θ_x can take any value. This singularity of the spherical coordinates has some practical consequences when it comes to define a covariance on the whole sphere, as we will see in Section 3.3.

Given a Gaussian random field $X_{x \in \mathbb{S}^2}$ and $x, y \in \mathbb{S}^2$, we will write $K_X(x, y)$ or $K_X(\theta_x, \theta_y, \varphi_x, \varphi_y)$ for its covariance function.

As in the Euclidean case, a simplifying assumption is to consider stationary models. A Gaussian process on the sphere is called *isotropic* if $K_X(x, y) = F(d(x, y))$, where $d(x, y)$ is the great circle distance on the sphere, given by the expression:

$$d(x, y) = \cos^{-1} (\sin \varphi_1 \cdot \sin \varphi_2 + \cos \varphi_1 \cdot \cos \varphi_2 \cdot \cos (\theta_2 - \theta_1)). \quad (4)$$

We refer to Gneiting[4], Porcu et al. [13], Jeong et al. [9] and references within for a state of the art on isotropic Gaussian processes.

However at the global scale, climatic data exhibits important nonstationarity effects: in particular, it is clear that climatic variables are typically correlated at a shorter range in the latitudinal direction than in the longitudinal. Jones [10] proposes to address this issue and defines *axially symmetric* Gaussian processes, which are stationary only in longitude variable, that is to say their covariance verifies

$$K_X(x, y) = F(\theta_x - \theta_y \mod 2\pi, \varphi_x, \varphi_y). \quad (5)$$

Furthermore an axially symmetric Gaussian process is said to be *latitudinally reversible* if

$$F(\theta_x - \theta_y \mod 2\pi, \varphi_x, \varphi_y) = F(\theta_x - \theta_y \mod 2\pi, \varphi_y, \varphi_x). \quad (6)$$

3.3 A new class of axially symmetric covariances

In the Euclidean case, a natural idea to obtain anisotropic covariances is to consider the product of covariances with respect to different coordinates. For example considering for $x, y \in \mathbb{R}^2$,

$$K(x, y) = \sigma e^{-\frac{|x_1 - y_1|}{r_1}} \cdot e^{-\frac{|x_2 - y_2|}{r_2}} \quad (7)$$

with distinct scale parameters r_1 and r_2 yields a model that is correlated at a longer range in a chosen direction. On the sphere one might consider in a similar way products of covariances in latitude and longitude, such as for $x, y \in \mathbb{S}^2$,

$$K(x, y) = \sigma e^{-\frac{|\varphi_x - \varphi_y|}{r\varphi}} \cdot e^{-\frac{|\theta_x - \theta_y|}{r\theta}}, \quad (8)$$

but in this case, since the longitude coordinate can take any value at the poles of the sphere (see Section 3.2), the covariance is not well defined when x or y is one of the two poles. Moreover, there is no way to chose a value for K at these points to obtain a continuous kernel, since when x (or y) closes to one of the pole, the value taken by $K(x, y)$ depend on the way x approaches the pole.

Discontinuity of the covariance kernel yields Gaussian fields that are discontinuous in L^2 and whose realizations are not almost surely continuous. This is unwanted in most applications where variables exhibits a continuous behavior over the sphere, such as climatic applications. In the case of (8) we obtain an expected singular behavior at the poles.

To address this issue, we consider instead of (8) a covariance

$$K(x, y) = \sigma e^{-\frac{d(x, y)}{r_{iso}}} \cdot e^{-\frac{|\varphi_x - \varphi_y|}{r\varphi}}. \quad (9)$$

Notice that since $(x, y) \mapsto d(x, y)$ and $x \mapsto \varphi_x$ are continuous function on the sphere, this time K is continuous everywhere. The term $e^{-\frac{|\varphi_x - \varphi_y|}{r\varphi}}$ can be seen as an additional decorrelation to the isotropic covariance $\sigma e^{-\frac{d(x, y)}{r_{iso}}}$ in the latitudinal direction.

Extending this idea to kernels other than the exponential, we obtain the following class of covariances.

Theorem 1. *Let K_{iso} be an isotropic covariance on the sphere and K_φ be a covariance on $[-\pi, \pi]$.*

The kernel defined by

$$K(x, y) = K_{iso}(x, y) \cdot K_\varphi(\varphi_x, \varphi_y) \quad (10)$$

is a latitudinally reversible, axially symmetric covariance on the sphere.

Furthermore, if K_{iso} and K_φ are continuous, K is continuous on the whole sphere, and as such, a Gaussian field with covariance K is continuous in L^2 sense, and has almost surely continuous trajectories.

Proof. As product of two valid covariances, K is a valid covariance (see Schur product theorem in Zhang [16]). Axial symmetry of K is a direct consequence of the isotropy of K_{iso} and the fact that K_φ does not depend on θ_x and θ_y .

In contrast with Jun and Stein [11, 12], whose approach seems to give richer classes of covariances, allowing for complex interactions between variables, our method permits to easily modify the classical isotropic covariances to obtain axially symmetric analogues. Porcu et al. [14] attain a similar goal, but focusing on the variograms. In comparison with Huang et al. [8], our covariances present the advantage to be continuous on the whole sphere.

4 Perspectives: performances and properties of axially symmetric models

In an ongoing work, we assess the performances of our new class of axially symmetric covariances, compared to the propositions of Jun and Stein [11, 12], Huang [8], Porcu et al. [14]. We take advantage of the two data sources that were presented in Section 2 to assess their performances in prediction, training on real data and assessing prediction with a global covering.

We expect that the simple expression (10) of our covariances allows us to study their differentiability at the origin in every direction, allowing for Gaussian processes with homogeneous trajectory regularity, or at the contrary for models which exhibit a regularity that depends on the direction of the trajectory.

References

1. Berrisford, P., Dee, D., Poli, P., Brugge, R., Fielding, M., Fuentes, M., Kållberg, P., Kobayashi, S., Uppala, S., Simmons, A.: The era-interim archive version 2.0 (1), 23 (2011). URL <https://www.ecmwf.int/node/8174>
2. Cressie, N.: Statistics for spatial data. *Terra Nova* **4**(5), 613–617 (1992)
3. Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, I., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.N., Vitart, F.: The era-interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society* **137**(656), 553–597. DOI 10.1002/qj.828. URL <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.828>
4. Gneiting, T.: Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19**(4), 1327–1349 (2013). DOI 10.3150/12-BEJSP06. URL <https://projecteuclid.org/euclid.bj/1377612854>
5. Heaton, M.J., Datta, A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R.B., Hammerling, D., Katzfuss, M., et al.: A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics* pp. 1–28 (2018)
6. Hitczenko, M., Stein, M.L.: Some theory for anisotropic processes on the sphere. *Statistical Methodology* **9**(1), 211–227 (2012). DOI 10.1016/j.stamet.2011.01.010. URL <http://www.sciencedirect.com/science/article/pii/S157231271100013X>
7. Hortal, M., Simmons, A.J.: Use of reduced gaussian grids in spectral models. *Monthly Weather Review* **119**(4), 1057–1074 (1991). DOI 10.1175/1520-0493(1991)119<1057:UORGGL>2.0.CO;2. URL [https://doi.org/10.1175/1520-0493\(1991\)119<1057:UORGGL>2.0.CO;2](https://doi.org/10.1175/1520-0493(1991)119<1057:UORGGL>2.0.CO;2)
8. Huang, C., Zhang, H., Robeson, S.M.: A simplified representation of the covariance structure of axially symmetric processes on the sphere. *Statistics & Probability Letters* **82**(7), 1346–1351 (2012). DOI 10.1016/j.spl.2012.03.015. URL <http://www.sciencedirect.com/science/article/pii/S0167715212001083>
9. Jeong, J., Jun, M., Genton, M.G.: Spherical Process Models for Global Spatial Statistics. *Statistical Science* **32**(4), 501–513 (2017). DOI 10.1214/17-STS620. URL <https://projecteuclid.org/euclid.ss/1511838025>
10. Jones, R.H.: Stochastic Processes on a Sphere. *The Annals of Mathematical Statistics* **34**(1), 213–218 (1963). DOI 10.1214/aoms/1177704257. URL <https://projecteuclid.org/euclid.aoms/1177704257>
11. Jun, M., Stein, M.L.: An Approach to Producing Space–Time Covariance Functions on Spheres. *Technometrics* **49**(4), 468–479 (2007). DOI 10.1198/004017007000000155. URL <https://amstat.tandfonline.com/doi/abs/10.1198/004017007000000155>
12. Jun, M., Stein, M.L.: Nonstationary covariance models for global data. *The Annals of Applied Statistics* **2**(4), 1271–1289 (2008). DOI 10.1214/08-AOAS183. URL <https://projecteuclid.org/euclid.aoas/1231424210>
13. Porcu, E., Alegria, A., Furrer, R.: Modeling temporally evolving and spatially globally dependent data. *International Statistical Review* **86**(2), 344–377 (2018)
14. Porcu, E., Castruccio, S., Alegría, A., Crippa, P.: Axially symmetric models for global data: A journey between geostatistics and stochastic generators. *Environmetrics* **30**(1), e2555 (2019). DOI 10.1002/env.2555. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2555>
15. Stein, M.L.: Spatial variation of total column ozone on a global scale. *The Annals of Applied Statistics* **1**(1), 191–210 (2007). DOI 10.1214/07-AOAS106. URL <https://projecteuclid.org/euclid.aoas/1183143735>
16. Zhang, F. (ed.): The Schur Complement and Its Applications. *Numerical Methods and Algorithms*. Springer US (2005). URL <https://www.springer.com/fr/book/9780387242712>

Analysis of the financial performance in Italian football championship clubs via GEE and diagnostic measures

Analisi delle performance finanziaria delle squadre di calcio di serie A via GEE e misure di diagnostica

Maria Kelly Venezuela¹, Anna Crisci², Luigi D'Ambra², D'Ambra Antonello³

Abstract

Football is undoubtedly the most powerful and most popular sport in Italy, linking communities and stirring emotions. The main goal of any Football Championship club is to achieve sport results. The study of the relationship between sport and economic results attracts the interest of many scholars belonging to different disciplines. Very informative is considered the connection, over short or long periods of time, between the points in the championship and the resource allocation strategies. The aim of this paper is to give an interpretation of this last link using the Generalized Estimating Equation (GEE) for longitudinal data. Some diagnostic measures and simulate envelope for checking the adequacy of GEE method will be presented and used.

Abstract

Il calcio in Italia è un fenomeno sociale che coinvolge intere comunità e continua ad aumentare il suo valore sociale ed economico. Lo studio della relazione tra i risultati sportivi ed economici riscuote l'interesse di tantissimi studiosi appartenenti a diverse discipline. Particolarmente stimolante è risultato il dibattito che lega, per ciascuna squadra di calcio, i punti in classifica alle capacità imprenditoriali del management sportivo in termini di allocazione delle risorse finanziarie e sportive. Obiettivo del presente lavoro è quello di dare un contributo in termini di interpretazione di quest'ultimo legame attraverso l'utilizzo delle Equazioni di Stima Generalizzate (GEE) per dati longitudinali. Alcune misure diagnostiche e metodi grafici per testare l'adeguatezza del metodo GEE saranno illustrati e utilizzati.

Key words: Italian Football championship clubs, Generalized Estimating Equations, Diagnostic Measures, Simulated envelope

¹ Insper Institute of Education and Research, São Paulo, Brazil

² University of Naples Federico II, Department Economic, Management, Institutions

³ University of Campania "Luigi Vanvitelli", Department of Economics

1. Introduction

Football is undoubtedly the most powerful and most popular sport in Italy, linking communities and stirring emotions. The main goal of any Football Championship club is to achieve sport results. Nevertheless, football has also become one of the most profitable industries, with a significant economic impact in infrastructure development, sponsorships, TV rights and transfers of players. Very informative is considered the connection between the points in the championship and the resource allocation strategies.

The Generalized Estimating Equation (GEE) [5] methodology has been introduced to extend the application of generalized linear models to handle correlated data. For repeated measures, nowadays GEE represents a method based on a quasi-likelihood function and provides the population-averaged estimates of the parameters.

The aim of this paper is to give an interpretation of the link between the points in the championship and the resource allocation strategies using the GEE. In particular, we analyze the impact that some financial indicators have on points made by football teams participating in the series A championship (2010-2015), by GEE for count data.

2. Overview Generalized Estimating Equation method

Let $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$ be a vector of response values and let $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})'$ be a $T \times K$ matrix of covariates, with $\mathbf{x}_{it} = (x_{it1}, \dots, x_{itK})'$, $i = 1, 2, \dots, n$ and $t = 1, 2, \dots, T$. To simplify notation, let $t_i = T$ without loss of generality.

The expected value and variance of measurement y_{it} can be expressed using a generalized linear model:

$$E(y_{it}|\mathbf{x}_{it}) = \mu_{it}$$

Suppose that the regression model is $\eta_{it} = g(\mu_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta}$ where g is a link function and $\boldsymbol{\beta}$ is an unknown $K \times 1$ vector of regression coefficients. The $\text{Var}(y_{it}|\mathbf{x}_{it}) = v(\mu_{it})\phi$, where v is a known variance function of μ_{it} and ϕ is a scale parameter which may need to be estimated. Mostly, v and ϕ depend on the distributions of outcomes. The variance-covariance matrix for \mathbf{y}_i is noted by $\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{\frac{1}{2}}$, $\mathbf{A}_i = \text{diag}\{v(\mu_{i1}), \dots, v(\mu_{iT})\}$ and the so-called “working” correlation structure $\mathbf{R}_i(\boldsymbol{\alpha})$ describes the pattern of measures within the subjects, which is of size $T \times T$ and depends on a vector of association parameters denoted by $\boldsymbol{\alpha}$.

The parameters $\boldsymbol{\beta}$ are estimated by solving: $U(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{D}_i' [V(\hat{\boldsymbol{\alpha}})]^{-1} \mathbf{s}_i = 0$ where $\mathbf{s}_i = (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)$ with $\hat{\boldsymbol{\mu}}_i = (\mu_{i1}, \dots, \mu_{iT})'$ and $(\hat{\boldsymbol{\alpha}})$ is a consistent estimate of $\boldsymbol{\alpha}$ and $\mathbf{D}_i' = \mathbf{X}_i' \boldsymbol{\Lambda}_i$ and $\boldsymbol{\Lambda}_i = \text{diag}(\partial \mu_{i1} / \partial \eta_{i1}, \dots, \partial \mu_{iT} / \partial \eta_{iT})$. Under mild regularity conditions $\hat{\boldsymbol{\beta}}$ is asymptotically distributed with a mean $\boldsymbol{\beta}_0$ and covariance matrix estimated based on the sandwich estimator:

$$\hat{V}_i^R = \left(\sum_{i=1}^n \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \sum_{i=1}^n \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{s}_i \mathbf{s}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \left(\sum_{i=1}^n \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \quad (1)$$

3. Choice of best model: adjusted R_{adj}^2 and modified Mallows' Cp (\tilde{C}_p) based on Wald Statistics

In this section we discuss the use of the adjusted coefficient of determination R_{adj}^2 for the GEE and modified Mallows' Cp for the choice of one or the best subsets. We show that the adjusted \tilde{R}_{adj}^2 based on Wald Statistics is:

$$\tilde{R}_{adj}^2 = 1 - \frac{n-1}{(n-K-1) + \tilde{Q}}$$

where

$$\tilde{Q} = [\mathbf{C}\hat{\boldsymbol{\beta}}]' [\mathbf{C}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\mathbf{C}']^{-1} [\mathbf{C}\hat{\boldsymbol{\beta}}]$$

is the Wald statistics with the GEE robust covariance matrix estimated under the null and denoted by $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$ and \mathbf{C} is a $(K-1) \times K$ matrix with its first column having all 0s, and its last $(K-1)$ columns being the $(K-1)$ identity matrix.

We select the best subset model from among the 2^K models. In order to select the best model, we consider a modified Mallows' Cp (\tilde{C}_p) [4] in GEE that is closely related to \tilde{R}^2 , where:

$$\tilde{R}^2 = \frac{\tilde{Q}/(n-k-1)}{1 + \tilde{Q}/(n-k-1)}$$

later,

$$\tilde{C}_p = (n-K) \frac{1 - \tilde{R}_p^2}{1 - \tilde{R}_K^2} + 2p - n \quad \text{with } p \leq K$$

where \tilde{R}_p^2 is calculated by considering the Wald statistics with the GEE robust covariance matrix and p regressors, while \tilde{R}_K^2 is calculated by considering the Wald statistics with the GEE robust covariance matrix and the complete set of K regressors.

Finally, Cantoni *et al.* [1] have proposed an extension of the Mallows' Cp for GEE approach, by:

$$GC_p = WRSS_A - N + 2dfc$$

where $WRSS_A = \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)' \hat{\mathbf{A}}_i^{-1} (\mathbf{y}_i - \hat{\mathbf{y}}_i)$ and $dfc = \text{tr}(\mathbf{H}^{-1}\mathbf{Q})$, $\mathbf{H} = n^{-1} \sum_i \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i$ and $\mathbf{Q} = n^{-1} \sum_i \mathbf{D}_i' \mathbf{A}_i^{-1} \mathbf{D}_i$. The matrix $\hat{\mathbf{A}}_i^{-1}$ can be replaced by $\mathbf{R}(\boldsymbol{\alpha})$ in order to consider the within correlations.

4. Regression diagnostics and simulated envelope

Model checking is an important aspect of regression analysis with independent or dependent observations [9]. Unusual data may substantially alter the fit of the regression model, and regression diagnostics identify subjects which might influence the regression relation substantially. Therefore, GEE approach also needs diagnostic procedures for checking the model's adequacy and for detecting outliers and influential observations. Graphical diagnostic plots can be useful for detecting and examining anomalous features in the fit of a model to data.

Regression diagnostic techniques that are used in the linear model [2] or in GLM [3] have been generalized to GEE. Venezuela *et al.* [6] described measures of local influence for generalized estimating equations. Here, we extend the diagnostic measures based on Cook distance, leverage and standardized residuals, of the regression model in GEE approach. Moreover, in order to identify possible outlier observations in the dataset and examine the adequacy of the fitted model, a suggestion is to plot the l th-ordered absolute values of the Pearson standardized residuals calculated from the fitted model to the dataset.

5. Results

The data used for our case study was obtained from the financial statements filed by the Serie A football teams. The period of study concerned the championship from season 2010/2011 up to 2014/2015. The focus of the analysis is to verify the impact that some financial indicators have on the points achieved by football teams.

We consider the following independent variables: Wage(W), Depreciation Expense of multi-annual player contracts (DEM), Revenue net of player capital gain(RNC), Net equity(NE). In addition, we have considered, on the bases a bivariate descriptive analysis, also the square effect of DEM (DEM^2), given the non-linear relationship between Point and DEM. Finally, the interaction between DEM and NE ($DEM*NE$) also was considered. We consider the QIC criterion in order to select the best working correlation structure among three structures: independent, exchangeable and AR-1. The results showed values of the QIC for these correlation structures very similar to each other, even though the AR1 structure was slightly smaller, it generates some difficulties in the choice of the best working correlation. Later, we have carried out a descriptive analysis for the within-subject correlation of variable Point by Year. We can note that the correlations are decreasing by Year and this would lead to choice of the AR-1 working correlation. Now, let's begin the choice of the best subsets using the criteria described in Sect. 3. In particular we selected the best subset within 2^K models. The choice falls on Model describe in table1, whose \tilde{C}_p is close to the number of variables (5.34) and lower GC_p by Cantoni

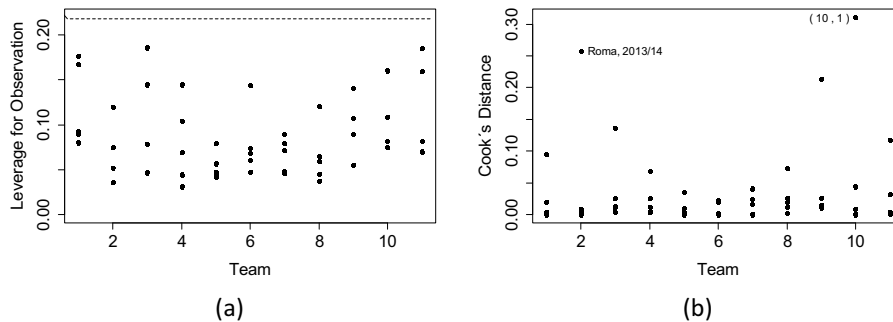
(35,67). Moreover, it shows the best R_{adj}^2 (0.55) and the lowest QIC (87.904). The Table 1 shows the output of this model. We have all significant variables, with the most important variable is DEM.

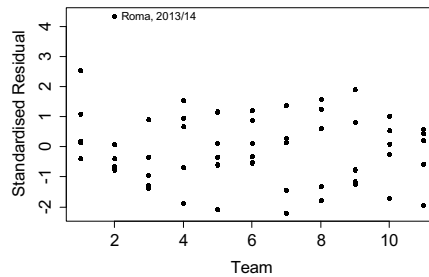
Table 1. Coefficient estimates of the Poisson regression model using AR(1) structure and with all observations.

Coefficients	Estimate	Robust S.E.	Robust z	p.value
Intercept	-21.069	11.611	-1.815	0.0696
DEM	3.074	1.330	2.311	0.0208
RNC	0.278	0.083	3.371	0.0007
NE	-0.718	0.196	-3.659	0.0003
DEM^2	-0.114	0.041	-2.804	0.0051
DEM*NE	0.045	0.012	3.836	0.0001
Correlation	0.347			

Wald statistic 71.08, Prob > Chi-square 0.000

Figure 1 shows the diagnostic measures using AR1 structure. In Figure 1(a), there is no observation playing leverage role in matrix of covariates. Figure 1(b) shows two observations as influence observations by Cook's distance, but one of them, that is the Roma team to the championship 2013/14, is also considered an outlier by the standardized residual in Figure 1(c).





(c)

Figure 1. Diagnostic measures for the Poisson regression model using AR(1) structure and with all observations.

The half-normal probability plot with simulated envelope (Figure 2) indicates a good fit with exception of one observation (Roma team to the championship 2013/14) that is outside the simulated envelope. Therefore, it can be concluded that the Poisson regression with AR1 correlation structure is adequate to explain the relation between the points achieved by football teams and financial variables.

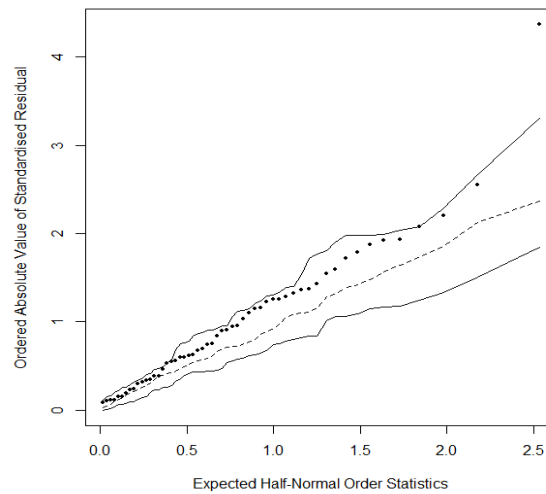


Figure 2. Half-normal probability plot with simulated envelope for the Poisson regression model using AR(1) structure and with all observations.

Just to be sure that, without the observation related to the Roma team for the

2013/14 championship, the fitted model would still be appropriate with the other observations, we consider a new estimation of the previous gee model with AR1 work correlation structure. Table 2 presents the new estimates of the coefficients described in the model, but without the observation related to Roma team to the championship 2013/14. Analyzing the results of this table, the same conclusion is maintained since all coefficients remain significant ($p.value < 0.05$). Then, the Roma team for the 2013/14 championship, there is not an influential point.

Table 2. Coefficient estimates of the Poisson regression model using AR(1) structure and without the observation related to Roma team to the championship 2013/14.

Coefficients	Estimate	Robust S.E.	Robust z	p.value
Intercept	-21.024	11.101	-1.894	0.058
DEM	2.967	1.294	2.293	0.022
RNC	-0.111	0.041	-2.744	0.006
NE	0.342	0.056	6.056	0.000
DEM^2	-0.711	0.208	-3.416	0.001
DEM*NE	0.044	0.012	3.562	0.000
Correlation	0.315			

References

1. Cantoni E, Flemming J, Ronchetti E. Variable selection for marginal longitudinal generalized linear models. *Biometrics* 2005;61(2):507–14.
2. Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**,15–18.
3. Cook, R. D and Thomas, W. (1989). Assessing influence on regression coefficients in generalized linear models. *Biometrika*, **76**, 741–750.
4. Crisci, A., D'Ambra, L. and Esposito, V. (2018). A Generalized Estimating Equation in Longitudinal Data to Determine an Efficiency Indicator for Football Teams. *Social Indicators Research*. <https://doi.org/10.1007/s11205-018-1891-6>
5. Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
6. Venezuela M.K., Botter D.A., Sandoval M.C. (2007) Diagnostic techniques in generalized estimating equations, *Journal of Statistical Computation and Simulation*, **77**:10, 879-888, DOI: 10.1080/10629360600780488

A statistical space-time functional model for air quality analysis and mapping

Un modello statistico spazio-tempo funzionale per l'analisi e la mappatura della qualità dell'aria

Yaqiong Wang, Alessandro Fassò and Francesco Finazzi

Abstract We introduce a functional space-time model for the analysis and mapping of air pollutants. The model is based on space-time latent variable and it is suitable for pollutant concentration data observed at high temporal frequency. The model is applied to ozone data in Beijing.

Abstract Introduciamo un modello spazio-tempo funzionale per l'analisi e la mappatura di inquinanti atmosferici. Il modello è basato su variabili spazio-temporali latenti ed è adatto per dati osservati ad alta frequenza temporale. Il modello è applicato a dati sull'ozono rilevati a Beijing.

Key words: EM algorithm, spline basis, high frequency data

1 Introduction and Data Description

Ground level ozone (O_3) plays a critical role in air quality and climate change. Also, ozone has adverse impact on human health [5]. Considering the increasing public concern on ozone, we attempt to estimate the effects from other pollutants and meteorological variables, and gain some insight into the diurnal cycle of O_3 , which peaks in the mid-day and reaches its minimum at night-time.

Yaqiong Wang

Peking University, 5 Yiheyuan Rd, Haidian Qu, Beijing Shi, China e-mail: yaqiongwang@pku.edu.cn

Alessandro Fassò

University of Bergamo, viale Marconi, 5 - 24044 Dalmine, Italy e-mail: alessandro.fasso@unibg.it

Francesco Finazzi

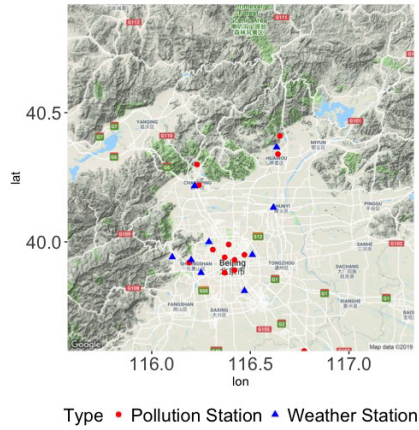
University of Bergamo, viale Marconi, 5 - 24044 Dalmine, Italy, e-mail: francesco.finazzi@unibg.it

In this study, we collect hourly concentrations of the ground ozone O_3 during year 2017 from twelve pollution monitoring stations in Beijing, which are directly managed by the Ministry of Environment and Protection (MEP). We also collect other four pollutant pollutants - particulate matters (PM_{10}), sulphur dioxide (SO_2), nitrogen dioxide (NO_2) and carbon monoxide (CO). All of the pollutants are measured in $\mu g/m^3$.

Secondly, we employ five hourly meteorological data - barometric pressure ($PRES$, in hectopascal), air temperature ($TEMP$, in degree Celsius), dew point temperature ($DEWP$, in degree Celsius), integrated rainfall ($IRAIN$, in millimetre), integrated wind speed (Iws , in meter per second) at nine weather stations from China Meteorological Administration (CMA). We match air quality stations and meteorological stations by their mutual distance. In addition to these meteorological variables, ultraviolet radiation is also a significant meteorological factor that influences O_3 generation. Therefore, we downloaded UVB (in J/m^2) with wavelengths between 200 and 440 nanometre from the European Centre for Medium-Range Weather Forecasts (ECMWF, <https://cds.climate.copernicus.eu>).

Figure 1 displays the spatial locations of the air quality stations with red dots as well as the meteorological stations with blue triangles [3].

Fig. 1 Red dots: the twelve air quality monitoring stations. Blue triangles: the nine meteorological stations.



2 Functional hidden dynamic geostatistical model

Let $y(s, h, t)$ be the profile observation at day t and hour h from site s with coordinates $s = (lat, lon)$. The loading coefficient $\mathbf{X}(s, h, t)$ contains the b -dimensional vector of covariates. $\Phi(h)$ is a set of p basis functions and $\mathbf{z}(s, t)$ is a p -variate latent spatio-temporal stochastic process. In this way, by describing random functional objects as linear combinations of the basis functions $\Phi(h)$ with Gaussian random

coefficients $\mathbf{z}(\mathbf{s}, t)$, we generalize the hidden dynamic geostatistical model [1] to functional data on the sphere. By further assuming that the $\boldsymbol{\beta}$ parameter depends on hour h , and that the $\varepsilon(\mathbf{s}, h, t)$ is a Gaussian measurement error with h -varying variance $\sigma_\varepsilon^2(h)$, the functional hidden dynamic geostatistical model can be extended as follows,

$$y(\mathbf{s}, h, t) = \mathbf{X}(\mathbf{s}, h, t)\boldsymbol{\beta}(h) + \boldsymbol{\Phi}(h)'\mathbf{z}(\mathbf{s}, t) + \varepsilon(\mathbf{s}, h, t), \quad (1)$$

$$\mathbf{z}(\mathbf{s}, t) = \mathbf{G}\mathbf{z}(\mathbf{s}, t-1) + \boldsymbol{\eta}(\mathbf{s}, t). \quad (2)$$

The dynamic of $\mathbf{z}(\mathbf{s}, t)$ is ruled by the $p \times p$ diagonal matrix $\mathbf{G} = \text{diag}(g_1, \dots, g_p)$, where the *diag* operation means the diagonal entries of \mathbf{G} are g_1, \dots, g_p . The persistency coefficient g_i is for the i -th variate of $\mathbf{z}(\mathbf{s}, t)$. The Gaussian innovation random field $\boldsymbol{\eta}(\mathbf{s}, t)$ is independent in time, with matrix spatial covariance function given by

$$\boldsymbol{\Gamma}(\mathbf{s}, \mathbf{s}') = \text{diag}(\tau_1 \rho(d(\mathbf{s}, \mathbf{s}'), \theta_1), \dots, \tau_p \rho(d(\mathbf{s}, \mathbf{s}'), \theta_p)).$$

Hence $\mathbf{V} = \text{diag}(\tau_1, \dots, \tau_p)$ is the variance-covariance matrix of $\boldsymbol{\eta}(\mathbf{s}, t)$. Moreover $\rho(d(\mathbf{s}, \mathbf{s}'), \theta)$ is the spatial correlation function with range parameter θ and geodesic distance $d(\mathbf{s}, \mathbf{s}')$ between \mathbf{s} and $\mathbf{s}' \in \mathbb{S}^2$.

We obtain the estimation of $\boldsymbol{\beta}(h) = (\beta_1(h), \dots, \beta_b(h))$ by modelling the elements as a linear combination of basis functions, that is $\beta_i(h) = \boldsymbol{\Phi}_\beta(h)'\mathbf{c}_{\beta,i}, i = 1, \dots, b$. In the same way, $\sigma_\varepsilon^2(h)$ can also be estimated by setting $\log(\sigma_\varepsilon^2(h)) = \boldsymbol{\Phi}_\varepsilon(h)'\mathbf{c}_\varepsilon$, where the logarithm transform is used to get the positive estimation $\sigma_\varepsilon^2(h)$. Thus, the unknown parameter set is $\Psi = \{\mathbf{c}_\beta, \mathbf{c}_\varepsilon, \mathbf{g}, \mathbf{v}, \boldsymbol{\theta}\}$, where $\mathbf{c}_\beta = (\mathbf{c}'_{\beta,1}, \dots, \mathbf{c}'_{\beta,b})$, $\mathbf{g} = (g_1, \dots, g_p)$, $\mathbf{v} = (\tau_1, \dots, \tau_p)$, and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$.

Appropriate numbers of basis functions should be chosen to estimate $\boldsymbol{\beta}(h), \mathbf{z}(\mathbf{s}, t)$, and $\sigma_\varepsilon^2(h)$. Here we choose the number of basis functions for $\boldsymbol{\beta}(h), \mathbf{z}(\mathbf{s}, t), \sigma_\varepsilon^2(h)$ to be 5, 7 and 5 respectively, which is reasonable based on our empirical knowledge. Additionally, since O_3 tends to be periodic, Fourier basis functions are chosen to analyze ozone data.

In the following context, we show how to select model covariate using the Akaike information criterion (AIC), and analyze the model estimation results based on the optimal subset of covariates.

3 Results

3.1 Covariates selection results

Table 1 displays the results of variable forward selection based on Akaike information criterion, which means starting with no covariates, and iteratively adding the most contributive covariates. For instance, at the beginning (Iter 0), we select the variable NO_2 , which results in the best performing model with maximum AIC. Then at the next iteration (Iter 1), we add one more covariate in previous model,

and the variable temperature (*TEMP*) is further selected. Moreover, Figure 2 shows the maximum AIC at each iteration. The improvement of model AIC is no longer significant after five iteration, therefore, we find the best subset of variables - PM_{10} , NO_2 , CO , $TEMP$, $DEWP$. Hence, the measurement equation for ozone data is,

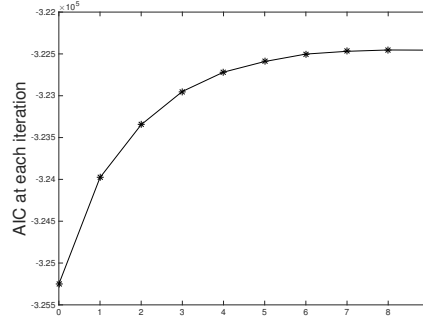
$$O_3(s, h, t) = \beta_{cons}(h) + x_{PM10}(s, h, t)\beta_{PM10}(h) + x_{NO2}\beta_{NO2}(h) + x_{CO}\beta_{CO}(h) + x_{TEMP}\beta_{TEMP}(h) + x_{DEWP}\beta_{DEWP}(h) + \Phi(h)'z(s, t) + \varepsilon(s, h, t), \quad (3)$$

where data are available at $h = 0, \dots, 23$, $s \in \{s_1, \dots, s_{12}\}$, and $t = 1, \dots, 365$.

Table 1 The corresponding AIC for the added variable at each iteration. ‘Selected’ means that the variables are already added at current iteration.

	Iter 0	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5	Iter 6	Iter 7	Iter 8	Iter 9
PM_{10}	-341133	-324685	-323346	Selected	Selected	Selected	Selected	Selected	Selected	Selected
SO_2	-341778	-325068	-323792	-323221	-322825	-322626	-322502	Selected	Selected	Selected
NO_2	-325252	Selected	Selected	Selected	Selected	Selected	Selected	Selected	Selected	Selected
CO	-340793	-324924	-323594	-323124	-322718	Selected	Selected	Selected	Selected	Selected
$PRES$	-342117	-324992	-323901	-323265	-322920	-322692	-322574	-322484	-322453	Selected
$TEMP$	-340039	-323978	Selected	Selected	Selected	Selected	Selected	Selected	Selected	Selected
$DEWP$	-341676	-325085	-323458	-322949	Selected	Selected	Selected	Selected	Selected	Selected
$IRAIN$	-341810	-325232	-323946	-323312	-322917	-322686	-322558	-322469	Selected	Selected
Iws	-341621	-325241	-323978	-323346	-322949	-322719	-322595	-322504	-322472	-322455
uvb	-341484	-325123	-323838	-323202	-322818	-322591	Selected	Selected	Selected	Selected

Fig. 2 Maximun AIC at each iteration. The improvement of AIC is no longer significant after five iteration. In this case, the best subset of variables - PM_{10} , NO_2 , CO , $TEMP$, $DEWP$ are selected.



3.2 Model estimation results

Figure 3 shows the estimated $\beta(h)$ and $\sigma_\epsilon^2(h)$ for model equation (3). Thanks to Fourier basis functions, the estimation result at the end of the day matches the beginning of the next day.

Among the three pollutants, PM_{10} and CO both have positive association with O_3 . Furthermore, the positive relation becomes stronger in the afternoon for PM_{10} while at noon for CO pollution. However, owing to the chemical coupling of O_3 and NO_2 , NO_2 has negative impact on O_3 , which is consistent with results in many studies [2]. For meteorological variables, the effect from temperature ($TEMP$) is positive, while the influence of dew point is not significant in the early morning. The results of χ^2 tests for the significance of covariates are reported in the Table 2 and indicate that all fixed effects are highly significant overall, even though the confidence bands of dew point variable contain zero in the early morning.

Since $\sigma_\epsilon^2(h)$ represents the unexplained portion of O_3 variance, the plot of $\sigma_\epsilon^2(h)$ shows that the model has more power in explaining O_3 during the day, especially from 5 a.m. to 5 p.m. [4].

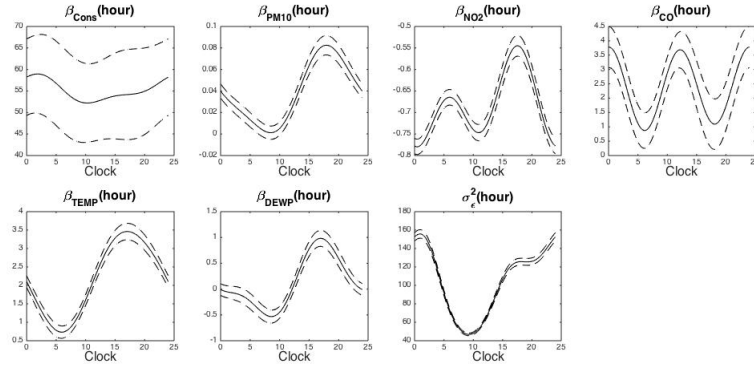


Fig. 3 Estimated $\beta_{cons}(hour)$, $\beta_{PM10}(hour)$, $\beta_{NO2}(hour)$, $\beta_{CO}(hour)$, $\beta_{TEMP}(hour)$, $\beta_{DEWP}(hour)$, and $\sigma_\epsilon^2(hour)$, with the 3σ bands.

Finally, in Table 3, we show the estimation and standard deviation of parameters relevant to the spatio-temporal process $\mathbf{z}(s, t)$, which are \mathbf{g} , $\boldsymbol{\theta}$ and \mathbf{v} . Most elements of \mathbf{g} are positive, within absolute value one, implying the stability of the 7-variate spatio-temporal latent process $\mathbf{z}(s, t)$. Compared with the geodesic distance in Beijing City (around 50 km), it makes sense that the value of $\boldsymbol{\theta}$ ranges from 12.60 km to 51.35 km. The average \mathbf{v} estimate is 2027, showing that the latent variable $\mathbf{z}(s, t)$ accounts for a large proportion of the O_3 variance.

Table 2 χ^2 tests for significance of fixed effects.

Covariate	χ^2 statistic	P-Value
Cons	398.39	0
PM10	1181.27	0
NO2	44785.52	0
CO	588.72	0
TEMP	3696.83	0
DEWP	511.70	0

Table 3 The estimated parameters \mathbf{g} , \mathbf{v} , $\boldsymbol{\theta}$ and their standard deviations.

Basis	\mathbf{g} transition		$\boldsymbol{\theta}$ elements		\mathbf{v} matrix	
	Value	Std	Value [km]	Std [km]	Value	Std
basis1	0.760	0.013	26.26	1.08	7090.92	237.10
basis2	0.325	0.021	27.78	0.42	4271.85	161.25
basis3	0.211	0.023	12.60	0.15	1196.27	50.18
basis4	0.121	0.024	13.90	0.13	747.67	34.02
basis5	0.127	0.025	15.93	0.025	570.37	26.99
basis6	-0.002	0.037	42.14	0.72	174.09	12.54
basis7	0.022	0.041	51.35	1.41	143.02	11.23

References

1. Calculi C, Fassò A, Finazzi F, Pollice A, Turnone A.: Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy. *Environmetrics* (2015), **26**(6), 406–417.
2. Chou CCK, Liu SC, Lin CY, Hsu CJ, Chang KH: The trend of surface ozone in Taipei, Taiwan, and its causes: Implications for ozone control strategies. *Atmospheric Environment*, 2006, **40**(21): 3898–3908.
3. Kahle D, Wickham H: ggmap: Spatial Visualization with ggplot2. *The R Journal* (2013), **5**(1), 144–161. <http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf>
4. Dohan J, Masschelein W.: The photochemical generation of ozone: Present state-of-the-art. *Ozone Science & Engineering* (1987), **9**(4), 315–334.
5. WHO: Air quality guidelines for Europe. World Health Organization (2002).

Tempering and computational efficiency of Bayesian variable selection

Tempering e l'efficienza computazionale della selezione bayesiana delle variabili

Giacomo Zanella and Gareth O. Roberts

Abstract We consider the problem of sampling from posterior distributions arising in the context of Bayesian variable selection models. We compare the importance tempering approach proposed in Zanella and Roberts (2019) with classical add-delete-swap Metropolis-Hastings schemes. Our results support the finding of Zanella and Roberts (2019) and suggest that importance tempering, and in particular the Weighted Tempered Gibbs Sampling algorithm, is a promising computational approach to achieve scalability for Bayesian variable selection.

Abstract *In questo articolo consideriamo il problema di simulare campioni dalla distribuzione a posteriori nel contesto di modelli bayesiani di selezione delle variabili. In particolare paragoniamo l'approccio importance tempering proposto in Zanella and Roberts (2019) con algoritmi classici di tipo Metropolis-Hastings. I nostri risultati supportano le conclusioni di Zanella and Roberts (2019) e suggeriscono che l'approccio importance tempering, e in particolare l'algoritmo Weighted Tempered Gibbs Sampling, sia un approccio promettente al fine di ottenere efficienza computazionale per modelli bayesiani di selezione delle variabili.*

Key words: Monte Carlo, MCMC, Importance Sampling, Bayesian Computation, Spike and slab Bayesian Variable Selection

Giacomo Zanella

Department of Decision Sciences, BIDS and IGIER, Bocconi University, via Roentgen 1, 20136 Milan, Italy. e-mail: giacomo.zanella@unibocconi.it

G.O. Roberts

Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK. e-mail: gareth.o.roberts@warwick.ac.uk

1 Introduction

Bayesian variable selection (BVS) models are a popular and successful approach to perform statistical inferences and prediction in large- p -small- n regression contexts, as well as to provide rigorous and interpretable uncertainty quantification regarding which variables should be included in the regression model (Chipman et al, 2001). However, gold-standard BVS models based on discrete priors run into computational problems when the number of regressors p is large. Recently, Zanella and Roberts (2019) proposed an importance tempering approach that achieves scalability to large p scenarios while still providing full Bayesian inferences and uncertainty quantification. In the simulation studies of Zanella and Roberts (2019), however, importance tempering was not compared to a classical and important class of sampling algorithms for BVS models (namely add-delete-swap Metropolis-Hastings schemes), which in principle could provide a scalable alternative to importance tempering (Yang et al, 2016). In this article we fill in this gap comparing importance tempering to add-delete-swap schemes, as well as other sampling algorithms, on various real datasets. Our results support the findings of Zanella and Roberts (2019) and suggest that importance tempering is a promising computational approach to achieve scalability in the context of Bayesian variable selection models with discrete priors.

The structure of the paper is as follows. In Section 2 we review the importance tempering approach describing the Tempered Gibbs Sampling algorithm and its weighted version in the context of generic target distributions. In Section 3 we describe the Bayesian variable selection models under consideration in this article. Finally in Sections 4 and 5 we list the sampling algorithms under consideration and compare their computational efficiency through a simulations study on real datasets coming from genomics.

2 Tempered Gibbs Sampling

Consider the problem of sampling from a probability density function $f(x)$ defined on a space \mathcal{X} , where x has d coordinates $x = (x_1, \dots, x_d) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_d = \mathcal{X}$. The Gibbs Sampler (GS) is a classical and widely-used algorithm to perform such task, which is based on the full conditionals of the target, denoted as $\{f(x_i|x_{-i})\}_{i \in \{1, \dots, d\}, x_{-i} \in \mathcal{X}_{-i}}$ where $f(x_i|x_{-i})$ is the conditional distribution of x_i given all the other coordinates $x_{-i} = (x_j)_{j \neq i}$ and $\mathcal{X}_{-i} = \times_{j \neq i} \mathcal{X}_j$. The Tempered Gibbs Sampler (TGS), instead, requires the specification of modified full conditionals, denoted as $\{g(x_i|x_{-i})\}_{i \in \{1, \dots, d\}, x_{-i} \in \mathcal{X}_{-i}}$. For each $i \in \{1, \dots, d\}$ and $x_{-i} \in \mathcal{X}_{-i}$, the function $x_i \mapsto g(x_i|x_{-i})$ needs to be a probability density function on \mathcal{X}_i absolutely continuous with respect to $f(x_i|x_{-i})$. It is not required that $\{g(x_i|x_{-i})\}_{i \in \{1, \dots, d\}, x_{-i} \in \mathcal{X}_{-i}}$ arise as the full conditional distribution of a global distribution $g(x)$ on \mathcal{X} . Given

$$p_i(x) = \frac{g(x_i|x_{-i})}{f(x_i|x_{-i})} \quad \text{for } i = 1, \dots, d; \quad Z(x) = \frac{1}{d} \sum_{i=1}^d p_i(x), \quad (1)$$

the TGS algorithm is defined as follows.

Algorithm TGS *At each iteration of the Markov chain do:*

1. (Coordinate selection) *Sample i from $\{1, \dots, d\}$ proportionally to $p_i(x)$.*
2. (Tempered update) *Sample $x_i \sim g(x_i|x_{-i})$.*
3. (Importance weighting) *Assign to the new state x a weight $w(x) = Z(x)^{-1}$.*

Steps 1 and 2 of the algorithm induce a Markov chain $x^{(1)}, x^{(2)}, \dots$ that is reversible with respect to the probability density function fZ on \mathcal{X} , defined as $(fZ)(x) = f(x)Z(x)$. Thus the weights in Step 3, $w(x) = Z(x)^{-1}$, are importance weights that correct for the difference between the target distribution f and the invariant distribution of the Markov chain, fZ . TGS is related to importance tempering (Gramacy et al, 2010), where Markov chain Monte Carlo (MCMC) is combined with simulated tempering and importance sampling. As shown in Proposition 1 of Zanella and Roberts (2019), under mild assumptions on f and $\{g(x_i|x_{-i})\}_{i \in \{1, \dots, d\}, x_{-i} \in \mathcal{X}_{-i}}$ the resulting Monte Carlo estimators defined as

$$\hat{h}_n^{TGS} = \frac{\sum_{t=1}^n w(x^{(t)})h(x^{(t)})}{\sum_{t=1}^n w(x^{(t)})} \quad (2)$$

are consistent estimators of posterior expectations in the sense that $\lim_{n \rightarrow \infty} \hat{h}_n^{TGS} = \mathbb{E}_f[h]$ almost surely for every function $h : \mathcal{X} \rightarrow \mathbb{R}$ with finite expectation $\mathbb{E}_f[h] = \int_{\mathcal{X}} h(x)f(x)dx < \infty$.

The TGS algorithm allows to replace the original full conditionals $\{f(x_i|x_{-i})\}_{i, x_{-i}}$ with the modified ones $\{g(x_i|x_{-i})\}_{i, x_{-i}}$ and corrects for the discrepancy with the importance weights in Step 3. Crucially, the importance sampling procedure is robust to high-dimensionality, in the sense that the variability of the weights $w(x)$ does not explode as d increases (unlike classical importance tempering). By choosing $\{g(x_i|x_{-i})\}_{i, x_{-i}}$ appropriately, TGS can avoid the slow mixing typical of GS in the presence of strong correlation among components of the target distribution (see Sections 2 and 3 of Zanella and Roberts (2019) for more details).

It can be shown that the TGS algorithm updates each coordinate with the same frequency, meaning that for every $i, j \in \{1, \dots, d\}$, x_i and x_j get updated the same number of times on average (see Remark 2 of Zanella and Roberts (2019)). In some contexts, however, it can be useful to have additional flexibility and direct control on the frequency of updating of each coordinate. The weighted Tempered Gibbs Sampling algorithm (WTGS) achieves this aim by introducing a weight functions $\eta_i : \mathcal{X}_{-i} \rightarrow (0, \infty)$ for each $i = 1, \dots, d$. Such functions are multiplied to $\frac{g(x_i|x_{-i})}{f(x_i|x_{-i})}$ in the selection probabilities $p_i(x)$ obtaining the following algorithm.

Algorithm WTGS *At each iteration of the Markov chain do:*

1. *Sample i from $\{1, \dots, d\}$ proportionally to*

$$p_i(x) = \eta_i(x_{-i}) \frac{g(x_i|x_{-i})}{f(x_i|x_{-i})},$$

2. *Sample* $x_i \sim g(x_i|x_{-i})$,
3. *Weight the new state* x with a weight $Z(x)^{-1}$ where $Z(x) = \zeta^{-1} \sum_{i=1}^d p_i(x)$ and $\zeta = \sum_{i=1}^d \mathbb{E}_{x \sim f}[\eta_i(x_{-i})]$.

The only difference between the WTGS and the TGS algorithm is in the different expression for the selection probabilities $p_i(x)$. TGS is as a special case of WTGS where $\eta_i(x_{-i}) = 1$. As shown in Proposition 6 of Zanella and Roberts (2019) the Markov chain induced by Steps 1 and 2 of the WTGS algorithm is still fZ invariant (where Z is defined in terms of the new selection probabilities p_i as in Step 3 of WTGS) and thus the Monte Carlo estimators \hat{h}_n^{WTGS} , defined analogously to (2), can again be used to approximate posterior expectations of interest $\mathbb{E}_f[h]$. The main difference with the TGS algorithm is that now the frequency of updating the i -th coordinate is proportional to $\mathbb{E}_{x \sim f}[\eta_i(x_{-i})]$. Thus, the weight functions $\eta_i(x_{-i})$ give control on the frequency of update of each coordinate.

3 Bayesian variable selection with discrete priors

Bayesian variable selection (BVS) models provide an elegant and effective approach to select a subset of explanatory variables in regression contexts and quantify the corresponding uncertainty (Chipman et al, 2001). Consider the standard Bayesian linear regression model

$$Y|\beta, \sigma^2 \sim N(X\beta, \sigma^2 \mathbb{I}_n), \quad \beta|\sigma^2 \sim N(0, \sigma^2 \Sigma), \quad p(\sigma^2) \propto \frac{1}{\sigma^2},$$

where Y is a $n \times 1$ response vector, X is an $n \times p$ design matrix, β is an $n \times 1$ vector of unknown regression coefficients, \mathbb{I}_n is the $n \times n$ diagonal matrix, $\sigma > 0$ is a unknown scalar standard deviation and Σ is the $p \times p$ prior covariance matrix for the p regression coefficients. In this context, the BVS approach is to introduce a vector of binary variables $\gamma = (\gamma_1, \dots, \gamma_p) \in \{0, 1\}^p$ indicating which regressor is included in the model and which one is not as follows: $\gamma_i = 1$ if the i -th regressor is included in the model and $\gamma_i = 0$ if it is excluded. For each $\gamma \in \{0, 1\}^p$ we obtain the following regression model

$$Y|\beta_\gamma, \gamma, \sigma^2 \sim N(X_\gamma \beta_\gamma, \sigma^2 \mathbb{I}_n), \quad \beta_\gamma|\gamma, \sigma^2 \sim N(0, \sigma^2 \Sigma_\gamma), \quad p(\sigma^2) \propto \frac{1}{\sigma^2},$$

where X_γ is the $n \times |\gamma|$ matrix containing only the included columns of the $n \times p$ design matrix X , β_γ is the $|\gamma| \times 1$ vector containing only the coefficients corresponding the selected regressors and Σ_γ is the $|\gamma| \times |\gamma|$ prior covariance matrix for the $|\gamma|$ selected regressors. Here $|\gamma| = \sum_{i=1}^p \gamma_i$ denotes the number of “active” regressors. The binary vector γ is then treated as an unknown parameter in a Bayesian framework

and assigned a prior distribution $p(\gamma)$ on $\{0, 1\}^p$. For example, a possible specification of $p(\gamma)$ is to assume

$$\gamma_i | h \stackrel{iid}{\sim} \text{Bern}(h) \quad i = 1, \dots, p,$$

where h is a prior inclusion probability, which can either be set to some fixed value in $(0, 1)$ or be given a prior distribution (e.g. a distribution belonging to the Beta family). The construction described above results in a joint distribution $p(Y, \beta_\gamma, \gamma, \sigma^2)$ over unknown parameters and data and a resulting posterior distribution $p(\beta_\gamma, \gamma, \sigma^2 | Y)$ for the unknown parameters.

BVS models can be computationally challenging due to the presence of the high-dimensional binary vector γ . Under standard prior specifications (such as the ones mentioned above) one can exploit conjugacy in linear regression models to integrate out analytically the unknown parameters β and σ , obtaining an explicit expression $p(\gamma | Y)$ for the marginal posterior distribution of γ (see e.g. Chipman et al, 2001). One is then left with the task of sampling from a distribution over $\{0, 1\}^p$, which is non-trivial due to the exponential size of such binary space. In the following sections we compare various algorithms to sample from $p(\gamma | Y)$, extending the simulation study of Zanella and Roberts (2019). Note that, given samples from $p(\gamma | Y)$, it is trivial to perform full posterior inferences on $p(\beta_\gamma, \gamma, \sigma^2 | Y)$ because $p(\beta_\gamma, \gamma, \sigma^2 | Y) = p(\beta_\gamma, \sigma^2 | \gamma, Y) p(\gamma | Y)$ and thus one simply needs to sample from $p(\beta_\gamma, \sigma^2 | \gamma, Y)$ using the conjugacy of standard linear regression models.

4 Sampling algorithms under comparison

In this article we compare four MCMC algorithms to sample from the distribution $p(\gamma | Y)$ described in Section 3. The standard way to draw samples from $p(\gamma | Y)$ is by performing Gibbs Sampling (GS) on the p components $(\gamma_1, \dots, \gamma_p)$, repeatedly choosing $i \in \{1, \dots, p\}$ either in a random or deterministic scan fashion and then updating $\gamma_i \sim p(\gamma_i | Y, \gamma_{-i})$. The second algorithm under comparison is the classical Add-Delete-Swap Metropolis-Hastings schemes (MH) (see e.g. Brown et al, 1998). Compared to GS, MH includes “swap” moves that can help exploring the posterior distribution in the presence of strongly correlated regressors. The third scheme under comparison is the Hamming Ball (HB) sampler recently proposed in Titsias and Yau (2017). Finally we consider the WTGS scheme proposed in Section 4.3 of Zanella and Roberts (2019), which is a specific implementation of the WTGS algorithm described in Section 2 for BVS models. For both the Gibbs Sampler (GS) and Weighted Tempered Gibbs Sampler (WTGS) we replace the Gibbs step with its Metropolised version, which is guaranteed to improve estimation accuracy for discrete target distributions (Liu, 1996). For WTGS we employ Rao-Blackwellization as described in Section 4.4 of Zanella and Roberts (2019). In the simulation study of Section 5 all algorithms are started from the empty model, i.e. $\gamma = (0, \dots, 0) \in \{0, 1\}^p$.

5 Simulations

We fit the BVS model of Section 3 to three real datasets, which we refer to as DLD data, TGFB172 data and TGFB data. The number of regressors and datapoints (n, p) for the three datasets are, respectively, $(192, 57)$, $(262, 172)$ and $(262, 10172)$. The DLD data, which is freely available from the supplementary material of Rossell and Rubio (2017), comes from a genomic study of Yuan et al (2016) based on RNA sequencing. See Rossell and Rubio (2017, Sec.6.5) for a short description of the dataset and the inferential questions of interest. The TGFB172 and TGFB datasets are human microarray gene expression data in colon cancer patients from Calon et al (2012). The TGFB172 data is obtained as a subset of the TGFB data as described in Rossell and Telesca (2017, Section 5.3). Both datasets are freely available from the supplementary material of Rossell and Telesca (2017, Section 5.3). In the BVS context, the main computational challenge comes from having a large number of regressors p . Thus we can roughly think at the three datasets as being of increasing computational difficulty.

We compare the four sampling algorithms described in Section 4 in terms of accuracy of estimation of marginal posterior inclusion probabilities (PIP), $\pi_i = p(\gamma_i = 1|Y)$ for $i = 1, \dots, p$. We employ the BVS model of Section 3, setting the prior covariance matrix to $\Sigma_\gamma = c(X_\gamma^T X_\gamma)^{-1}$ with $c = n$, which corresponds to the commonly used unit information g -prior, and setting the prior inclusion probability h to $5/p$ corresponding to a prior expected number of active regressors equal to 5 (we also experimented with other priors for h , such as a uniform prior on $[0, 1]$, obtaining similar results in terms of relative efficiency among algorithms).

We ran the WTGS algorithm for 1000 iterations for the two smaller datasets (DLD and TGFB172) and for 30000 iterations for the larger dataset (TGFB), discarding the first 10% of samples as burnin. The number of iterations of GS, MH and HB were chosen to have the same CPU runtime of WTGS. For each algorithm and each dataset under consideration, we perform 20 independent MCMC runs, obtaining 20 Monte Carlo estimates $(\hat{\pi}_i^{(j)})_{j=1}^{20}$ of each PIP π_i . We then compare PIP estimates obtain over different runs of the same algorithm, i.e. $\hat{\pi}_i^{(j)}$ and $\hat{\pi}_i^{(j')}$ for $j \neq j'$, in order to assess the Monte Carlo variability of the estimators produced by each algorithm. The results are displayed in Figure 1: each box corresponds to a combination of algorithm and dataset and the points plotted are $(\hat{\pi}_i^{(j)}, \hat{\pi}_i^{(j')})$ for $i \in \{1, \dots, p\}$ and $j, j' \in \{1, \dots, 20\}$ with $j \neq j'$. Thus, points close to the diagonal indicate algorithms that produce Monte Carlo estimators with small variance and allow to perform reliable inferences. Figure 1 clearly shows that WTGS produce Monte Carlo estimators with significantly smaller variance, especially for large values of p . Compared to the results displayed in Zanella and Roberts (2019, Sec.5.3), Figure 1 additionally shows that, in this context, MH performs only slightly better than GS and significantly worse than WTGS. This supports the results of Zanella and Roberts (2019) and further suggests that WTGS provides state of the art performances in the context of sampling algorithms for BVS models.

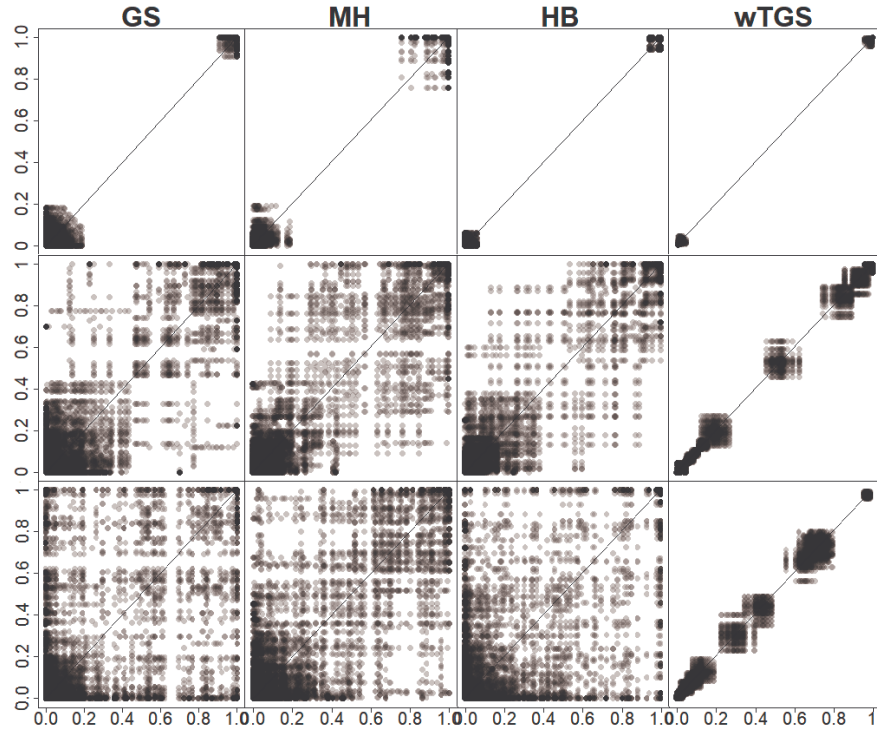


Fig. 1 Comparison of the variability of the Monte Carlo estimators produced by the sampling algorithms described in Section 4 (GS, MH, HB and WTGS on the first, second, third and fourth column respectively) on three real datasets (DLD, TGFB172 and TGFB data on the first, second and third row respectively). Points close to the diagonal lines indicate estimators with small variability, see Section 5 for more details.

Appendix

In this appendix we provide a compact and self-contained description of the Markov transition kernels associated to the four sampling algorithms under consideration in Section 4. Given the current state of the Markov chain $\gamma = (\gamma_1, \dots, \gamma_p) \in \{0, 1\}^p$ each algorithm generates the next state as follows:

- Metropolisised Gibbs Sampler (GS): sample i uniformly from $\{1, \dots, p\}$ and switch γ_i to $1 - \gamma_i$ with probability $\min\{1, \frac{p(1-\gamma_i|\gamma_{-i}, Y)}{p(\gamma_i|\gamma_{-i}, Y)}\}$.
- Add-Delete-Swap Metropolis-Hastings sampler (MH): with probability $\frac{1}{2}$ perform one GS iteration, otherwise do the following:
 1. sample i uniformly from $\{i \in \{1, \dots, p\} : \gamma_i = 1\}$ and j uniformly from $\{j \in \{1, \dots, p\} : \gamma_j = 0\}$
 2. Define $\gamma^* = (\gamma_1^*, \dots, \gamma_p^*)$ as $\gamma_i^* = 0$, $\gamma_j^* = 1$ and $\gamma_\ell^* = \gamma_\ell$ for $\ell \notin \{i, j\}$.

3. Move to γ^* with the corresponding Metropolis-Hastings probability.
- Hamming Ball sampler (HB)
 1. Sample j uniformly from $\{1, \dots, p\}$ and switch γ_j to $1 - \gamma_j$.
 2. Sample i from $\{1, \dots, p\}$ with probability proportional to $1 - p(\gamma_i | \gamma_{-i}, Y)$ and switch γ_i to $1 - \gamma_i$.
- Weighted Tempered Gibbs Sampler (WTGS)
 1. Sample i from $\{1, \dots, p\}$ with probability proportional to $p_i(\gamma) = \frac{p(\gamma_i=1|\gamma_{-i}, Y) + k/p}{2p(\gamma_i|\gamma_{-i}, Y)}$ for some fixed k (e.g. $k = 5$) and switch γ_i to $1 - \gamma_i$.
 2. Weight the new state γ with a weight $w(\gamma) = Z(\gamma)^{-1}$ where $Z(\gamma) \propto \sum_{i=1}^p p_i(\gamma)$.

References

- Brown PJ, Vannucci M, Fearn T (1998) Bayesian wavelength selection in multi-component analysis. *Journal of Chemometrics* 12(3):173–182
- Calon A, Espinet E, Palomo-Ponce S, Tauriello DV, Iglesias M, Céspedes MV, Sevillano M, Nadal C, Jung P, Zhang XHF, et al (2012) Dependency of colorectal cancer on a TGF- β -driven program in stromal cells for metastasis initiation. *Cancer cell* 22(5):571–584
- Chipman HA, George EI, McCulloch RE (2001) The Practical Implementation of Bayesian Model Selection. Institute of Mathematical Statistics Lecture Notes-Monograph Series 38:65
- Gramacy R, Samworth R, King R (2010) Importance tempering. *Statistics and Computing* 20(1):1–7
- Liu JS (1996) Peskun’s theorem and a modified discrete-state gibbs sampler. *Biometrika* 83(3)
- Rossell D, Rubio FJ (2017) Tractable bayesian variable selection: beyond normality. *Journal of the American Statistical Association* (just-accepted)
- Rossell D, Telesca D (2017) Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association* 112(517):254–265
- Titsias MK, Yau C (2017) The Hamming ball sampler. *Journal of the American Statistical Association* pp 1–14
- Yang Y, Wainwright MJ, Jordan MI (2016) On the computational complexity of high-dimensional bayesian variable selection. *The Annals of Statistics* 44(6):2497–2532
- Yuan T, Huang X, Woodcock M, Du M, Dittmar R, Wang Y, Tsai S, Kohli M, Boardman L, Patel T, et al (2016) Plasma extracellular rna profiles in healthy and cancer patients. *Scientific reports* 6:19,413
- Zanella G, Roberts G (2019) Scalable Importance Tempering and Bayesian Variable Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, in press

Dimensions and links for Hate Speech in the social media

Dimensioni e legami per i discorsi di odio nei social media

Emma Zavarrone, Guido Ferilli

Abstract The generation of the hate speech intangible sub-culture seen as an ultimate product of the synthesis of unstructured data is the context in which this analysis is born with reference to the words of hatred towards migrants. This paper is configured as a methodological advancement of a research European always on the hate speech and is characterized by the application of an unsupervised classification method (LDA) and the consequent search for links between the terms subject to LDA classification.

Abstract *La generazione di sotto cultura dai discorsi di odio, vista come un prodotto ultimo della sintesi di dati non strutturati, è il contesto in cui nasce questa analisi con riferimento alle parole di odio nei confronti dei migranti. Questo lavoro si configura come un avanzamento metodologico di una ricerca europea e sull'hate speech ormai conclusa ed è caratterizzato dall'applicazione di un metodo classificazione unsupervised (LDA) e della conseguente ricerca di legami tra i termini così classificati.*

Key words: LDA, eigenvector centrality, hate speech, bipartite network

1 Introduction

The iper-availability of unstructured data can be seen as the proxy of the wide cultural transformation in act, in which in principle the production of shared contents is accessible to everyone, to the extent to speak of a new

¹ Emma Zavarrone, emma.zavarrone@iulm.it; Università IULM,.; Guido Ferilli, Università IULM

paradigm of generation of information and therefore of knowledge. The concept of knowledge implies the aspiration towards more inclusive forms of evolution in which some manifestations of conflict shouldn't take place. Instead, we are witnessing a controversial process in which the listening of the social media in multiple circumstances gives the hate speakers relevant positions in terms of influencer. Hate speech is usually directed towards anything that channels the denigration of a specific group of people using uneducated or irrelevant language (Malmasi & Zampieri, 2018). To date, there is no common definition of hate speech, there is a propensity for the adoption of the definition used by the European Community (2005) in which *“the term ‘hate speech’ shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.”* In fact, over time hate speech has changed its medium used for expression, increasingly preferring to that of social media, therefore it is important to reflect also on the article 2.1 that take into account the Additional Protocol of the Convention on Cybercrime (28 January 2003)¹, which states that *“racist and xenophobic material” means any written material, any image or any other representation of ideas or theories, which advocates, promotes or incites hatred, discrimination or violence, against any individual or group of individuals, based on race, color, descent or national or ethnic origin, as well as religion if used as a pretext for any of these factors*”. The Italian regulatory system adheres to the European one, and there are several laws in force that protect the freedom of expression of the individual. It is not the purpose of this work to reflect on the boundaries between freedom of speech and hate speech, but there is still a tendency to use, on certain issues, terms closer to the tones of denigration. For example, monitoring conversations on different media (Facebook, Twitter, Instagram), over time, it emerges that with reference to the general themes such as migrants and women, the terms of hate are used and often these two themes are a driver to comment the political scene. This panorama has also encouraged reflections for the comparison on the theme of migrants in Italy, Greece, Great Britain, Croatia, Bulgaria, the Czech Republic and Romania, thus carrying out a two-year research (2017-2018) financed by the European Justice Programme². The structure of the research compared specific sets of online keywords for two weeks between May and June 2017

¹ <http://conventions.coe.int/Treaty/>

² Project “Coalition of Positive Messengers to Counter Online Hate Speech”, No JUST/2015/RRAC/AG/BEST/8931.

and two months in 2016 one month before and one month after a relevant event related to the migrants (and chosen by each project partner), using tweets, posts, blogs not only of the network but also of three online newspapers, chosen to represent the different social and political orientations in each country. The main objectives were to verify whether:

1. Hate speech online against immigrants and refugees has the same value in different countries;
2. If the presence of online hates narratives could be the latent engine to trigger social disorder.

The results of point 1 show that hate speech is regularly used when people are speaking of migrants in all the 7 countries, and that the nature of the words reflects the evolution of the historical context that has characterized the different countries. In addition, there is a time dependency correlation between the uses of the terms that has been estimated through the employment of an AutoCM, Artificial Neural Networks tool (Buscema, Ferilli, Massini & Zavarrone, 2018). In this work, the focus will be on the unsupervised classification of topics closer to the migration (A_1) and on the discovery of relationships between the words assigned to different topics (A_2). Section 2 deals out the data, Section 3 presents a brief recall on the applied methodologies and the fourth one will be devoted to the results. In the last section, some limits will be discussed.

2 Data

The design of the research is depicted in Fig. 1a where two social media, FB and Twitter, have been engaged. Each project 'partner has proposed ten keywords for denoting the hate speech on line oriented to the migrants. The generation of keywords was executed through the focus group in each country. The lists of keywords are twofold: one written in own language and another one in English, in order to have a common framework for the evaluation (Tab.1).

Table 1: Pattern of selected keywords and number of tweets for Twitter users in last row (2017)

UK	Bulgaria	Italy	Romania	Czech Rep	Croatia	Greece
invaders	мигрант, мигранти	pocket money	Islamiști	Uprchlíci	izbjeglička kriza	Τρομοκρατίας
fakefugees	бежанец, бежанци	prima gli italiani	Musulmani	Migranti	teroristi	Μουσουλμάνο ς
gimmigrants	заплаха	Non ci rubano il lavoro	Migrant/M igranți	Islám	kozojebi	Εγκληματίες
#NoRefugees	терорист, терористи	magrebini	Refugiați	Islamisté	ustaše	Ισλάμ
rapefugees	ислямист, ислямист	finti profughi	Teroriști	Terorismus	pederi	Λάθρο-
Vermin	нелегален, нелегални	immigrati	Cioară/Cio ri	Pražská kavárna	balije	Τζιχαντιστές
Foreigners	талибани	clandestini	Țigan	Vlastizrádci	četnici	Εισβολή
Romanians Deport	на сапун чернилка	islamici barconi	Poponar Jidan	Sluničkáři Pravdoláskáři	antife bodljikava žica	Βιάζουν Πρόσφυγες
Migrants	вън	ciabattanti	Bozgor	Multikultural ismus	izbjeglice	Μετανάστες
91.893	182	2.149	89	268	2,140	1.934

Table 2: Scheme of data collection

Year	Sources	Twitter	Facebook
2017 (28/05-10/06)	Users	7 corpus (each for language)	
	On-line newspapers	21 corpus (7 each for language*3 political positioning)	21 (7 each for language* 3 political positioning)
2016 (1/05-31/05)	On-line newspapers	21 corpus (7 each for language* 3 political positioning)	21 (7 each for language* 3 political positioning)

The body of techniques labelled under the information retrieval has allowed the identification and the selection of the data on Twitter and Facebook. Tweets posts and messages can be filtered by keywords, users, language, and location. These data become the raw data for developing the analysis. The

data collection has generated overall 91 corpus in different languages and collected from different sources and years as shown in Tab.2. Different levels both of sensitivity to the topic under exam and internet penetration rate in the countries partners, in addition with the habits of use of social media configure for the users only, seven samples of tweets with dissimilar size (last row of Tab.1) characterized by high/low lexical richness. In Romania, Czech Rep and Bulgaria, the use of Twitter was not widespread than Facebook (StatCounter, 2017), so it can explain the differences among the samples.

3 Methodologies

In the European project were applied textual, social network and neural network methodologies in order to satisfy the prefixed aims.

An intermediate result useful for developing the core of analysis in the European project has been represented by the Document Term Matrices (DTM).

In this paper with the formulation of two hypothesis, another piece of analysis has been added. From DTM we developed an unsupervised classification using Latent Dirichlet Allocation (LDA) approach (Blei and Jordan, 2003) for the A_1 and we apply the social network analysis on incidence matrix derived from LDA for A_2 . For LDA each document is as a mixture of topics, and each topic as a mixture of terms. The content of the documents tends to overlap each other in terms of content, using a scheme of posterior probability.

The literature proposes some methods, such as Relational Topic Model (Chang and Blei, 2013) based on the distance between their topic proportions, in order to obtain a textual network from LDA. In this paper, we propose to use the posterior probability as entries of an incidence matrix, (\mathbf{I}_{wdxt}) composed by words on the rows and topics on the columns. In the first stage, we analyse the two-mode network for discovering which words can be considered as links among the topics. In the second stage of the analysis our research focuses on co-affiliation of the terms and topics where the matrix \mathbf{I} , after some transformations, becomes the affiliation matrix (\mathbf{I}'_{wxt}) with each cell has been indicated the binary value, 1: if i -th terms belongs to the topic j -th; 0: otherwise. The \mathbf{I}' can be analysed with the conversion approach (Everett & Borgatti, 2013), so two matrices can be obtained: a) terms to terms matrix (**TE**) and topics to topics matrix (**TO**). Our attention is focused

on both matrices therefore to the measure of centrality. The eigenvector centrality (Bonacich, 1972) has been chosen for answering to the second research question.

4 Results

The UK tweets' corpus, because of its size, has been selected for testing the A_1 and A_2 . The corpus, after data-cleaning operations has been reduced in a document term matrix based on 91.893 rows (documents), and 225 columns (terms), so the matrix configures very sparse. Figure 2 shows the distribution of the 25 most frequent terms. At first glance, the terms seem refer to hybrid topics: Trump shares the content of tweets with Theresa May and hate. This situation brings us to apply LDA, (A_1), we have tested several models (Figure 3) and based on perplexity measure (Brown and Della Pietra, 1992), we chose LDA with three topics. The literature asserts some limits on this measure but for our aim it well describes the concurrent topics. The Figure 4 shows the classification of the first 10 terms for each topic which can be denominated: Topic1: Global, Topic2: Local, Topic 3: Hate Driven.

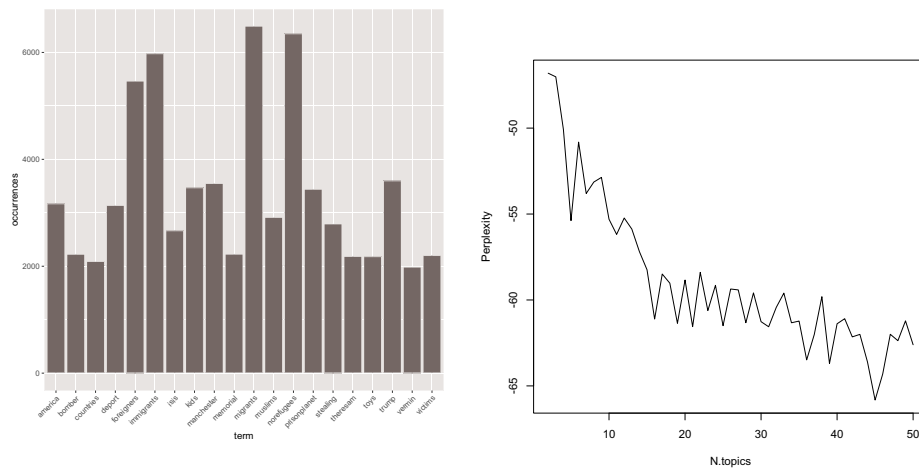
This classification highlights how a generic theme can produce different flow of information and contribute to the generation of intangible sub-culture.

A_2 is oriented to explore a system of relationship among the terms that belong to each topic since a single term can be in more topics. This characteristic becomes the innovative proposal of the paper: developing a network of terms, an intangible sub-culture, based on LDA approach.

The existence of a core of words, able to connect the topic, has been tested with the exploration of the network as shown in Fig. 4, where *foreigners*, *attack*, *people* are the central terms included in a triad. The terms belong to the triad offer a complex interpretation of the reality, the most important lecture could be: the wave of hate fueled by the fear towards the foreigners. The analysis of two networks, from $TE_{(225 \times 225)}$ and $TO_{(3 \times 3)}$ matrices, allows to capture some additional information. In detail, the high level of eigenvector centrality value (0.97) has been associated with the terms *foreigners*, *terror* and *women*, but a comparison on methods for entries dichotomization could be useful.

Figure 2: Most frequent terms

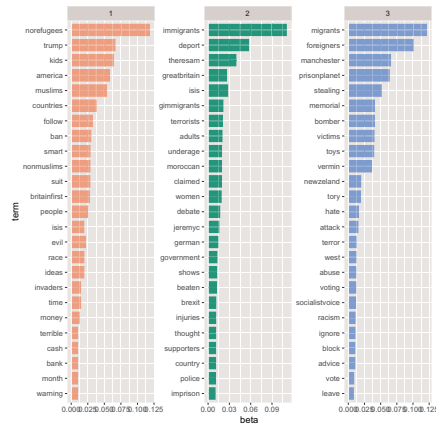
Figure 3: Distribution of perplexity



5 Discussion and further development

The most important result addressed to the presence of more latent dimensions or topics that governs the tweet's content though, with an initial selection process based on the specific keywords. The wide and intensive use of social media generates neighbor cultural contents. The first aspect to put in evidence is the context in which people use the social media: the data collection was in a specific time with terroristic attacks in UK and the referendum for the Brexit. In particular the time period choices in 2017 was one month before, and another after the referendum. In order to verify the robustness of the proposed approach, the analysis should be monitoring in constant way and extended to the rest of the countries. From the methodological perspective, the introduction of other measure for selecting the number of topics could be advisable. More care has to be devoted to sparse incidence matrix: alternative schemes recently added in the literature could be inserted and compared.

Figure 4: Distribution of the first 25 terms for three Topics and bipartite Network bipartite



References

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. The Journal of Machine Learning Research. 3, 993–1022 (2003).
2. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. Journal of mathematical sociology, 2(1), 113-120(1972).
3. Brown, P. F., Della Pietra, V. J. D., Mercer, R. L., Pietra, S. A. D., & Lai, J. C. :An estimate of an upper bound for the entropy of English. Computational Linguistics, 18(1), 31-40(1992).
4. Buscema, M., Ferilli, G., Massini, G. & Zavarrone, E.: Media content analysis on line hate speech, Project reference number: JUST/2015/PRAC/AG/BEST/8931 (2018)
5. Everett, M. G., & Borgatti, S. P.: (2013). The dual-projection approach for two-mode networks. Social Networks, 35(2), 204-210 (2013).
6. Malmasi, S. and Zampieri, M.: Challenges in discriminating profanity from hate speech. Journal of Experimental & Theoretical Artificial Intelligence 30.2, 187-202(2018).
7. <http://gs.statcounter.com>



Section 3. Contributed Papers

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

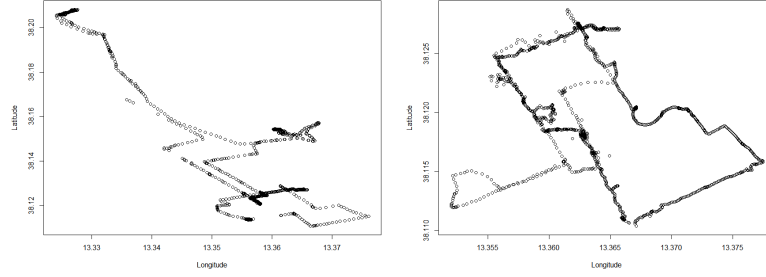


Fig. 1 Two illustrative units' behaviors.

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 ε -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 ε -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 ε 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 ε and $MinPts$ if there is a chain $p_1, \dots, p_l, p_l = q, p_l = p$ such that p_{l+1} is directly density-reachable from p_l . An object p is **density-connected** to object q with respect to ε and $MinPts$ if there is an object o such that both, p and q are density reachable from o with respect to ε 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 ε and $MinPts$, then $p \in C$;
- $\forall p, q \in C$: p is density-connected to q with respect to ε 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 ε and $MinPts$. If p is a core point, this procedure yields a cluster concerning ε 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.

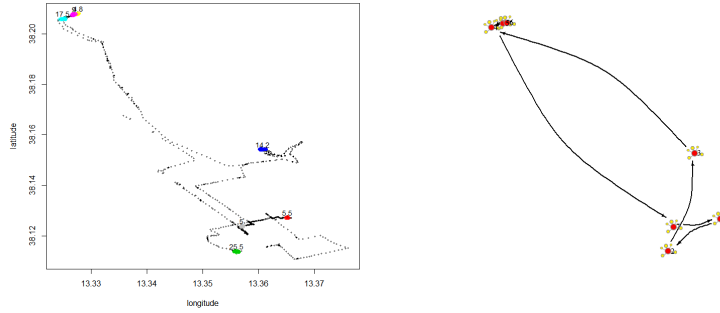


Fig. 2 Raw GPS data and clusters obtained with the DBSCAN algorithm (on the left); the numbers over each cluster represent the times spent by the unit into that cluster. Nodes and links (on the right) represent points of interest and sequence of visit, respectively, for the considered unit.

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 obtaining 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.

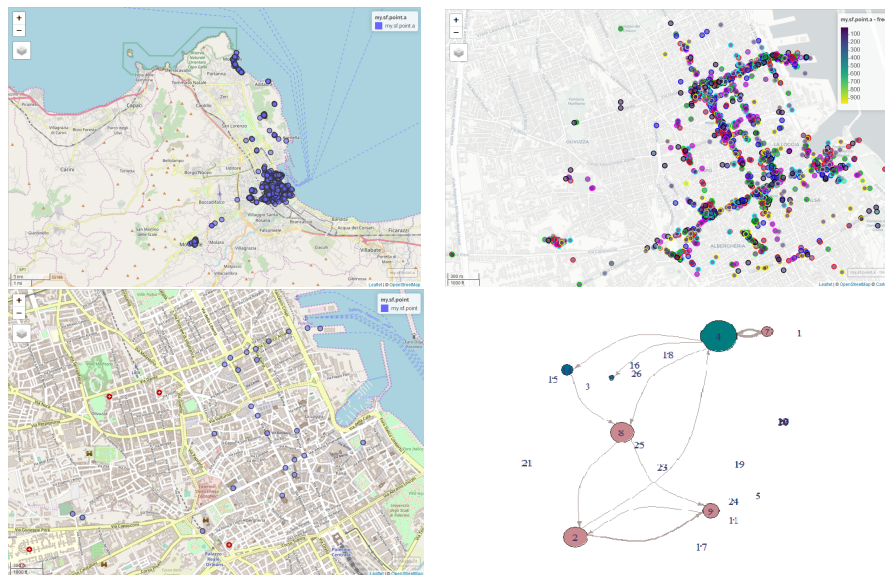


Fig. 3 Centroid Cluster recovered by using Algorithm 1. On the left side the cluster of the entire city of Palermo and on the right side the zoom on the city center.

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.

Local inference on functional data based on the control of the family-wise error rate

Inferenza locale per dati funzionali basata sul controllo del family-wise error rate

Konrad Abramowicz, Alessia Pini, Lina Schelin, Sara Sjöstedt de Luna, Aymeric Stamm, and Simone Vantini

Abstract In this work we focus on the problem of local inference for functional data. We describe a unified framework for testing hypotheses on functional data in a local perspective. The result of the testing procedures within the unified framework is an adjusted p -value function that can be used to select the areas of the domain responsible for the rejection of the null hypothesis. We discuss how different state-of-the-art inferential procedures fall within the framework, and briefly describe a novel testing procedure with sound theoretical properties.

Abstract *Questo lavoro si concentra sul tema dell'inferenza locale per dati funzionali. Viene descritto un framework unificato per effettuare test di ipotesi su dati funzionali con una prospettiva locale. Il risultato delle procedure che ricadono all'interno del framework è una funzione p -value aggiustata che può essere utilizzata per selezionare le aree del dominio responsabili del rifiuto dell'ipotesi nulla. Viene discusso come diverse procedure inferenziali proposte in letteratura rientrino nel framework proposto. Inoltre viene brevemente descritta una nuova procedura con valide proprietà teoriche.*

Konrad Abramowicz and Sara Sjöstedt de Luna
Department of Mathematics and Mathematical Statistics, Umeå University, 901 87 Umeå, Sweden
e-mail: konrad.abramowicz@umu.se, sara.sjostedt.de.luna@umu.se

Alessia Pini
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123, Milan, Italy e-mail: alessia.pini@unicatt.it

Lina Schelin
Unit of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, 901 87 Umeå, Sweden e-mail: lina.schelin@umu.se

Aymeric Stamm
UMR CNRS 6629, Laboratoire de Mathématiques Jean Leray, Nantes, France e-mail: Aymeric.Stamm@univ-nantes.fr

Simone Vantini
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milan, Italy e-mail: simone.vantini@polimi.it

Key words: nonparametric inference, functional data analysis, family-wise error rate

1 Introduction

Functional data analysis is the analysis of functional data, i.e., random curves embedded in infinite dimensional spaces. Functional data can be seen as an extreme case of multivariate data with infinite dimensionality (see for instance [3, 4, 8] and references therein); in this work we specifically focus on null hypothesis significance testing for functional data.

Just as inference for multivariate data can be performed either on a multivariate or on a component-wise perspective, inference for functional data can be approached by either a global or a local perspective. In the case of global inference, a single test is performed on the entire domain, and a p -value is computed, suggesting whether to reject or not the null hypothesis of the test on the entire domain (e.g., [5]). In the case of local inference instead, the null hypothesis is tested locally by defining a p -value function that assigns a p -value to each point of the domain (e.g., [7, 1]). Just as component-wise inference for multivariate data performs feature selection, local inference for functional data performs domain selection: the portions of the domain responsible for the rejection of the null hypothesis are naturally selected by assigning a threshold to the p -value function. Even though global inferential techniques are mathematically sound and powerful procedures, since they treat functional data as objects defined in infinite dimensional spaces, local inferential techniques are more and more required in the practice, since the information on the areas of the domain where the null hypothesis is rejected is often very important for practitioners.

In the local setting, assuming that functional data are smooth, or at least continuous, it is straightforward to compute a p -value at every point of the domain. Such p -values lead to a p -value function that controls only pointwise the probability of committing a type I error. For this reason, we refer to it as an *unadjusted* p -value function. One of the main issues in this framework is how to efficiently adjust such function, in order to provide an error control over the entire domain. In multivariate data analysis, several techniques have been proposed to adjust component-wise p -values to provide a control of the so called family-wise error rate (FWER), that is the probability of committing at least one type I error over all components. In this work we briefly discuss the extension of FWER to functional data, and describe a unified framework for controlling the functional FWER on domain subsets. An extensive description of the framework is reported in [2]. The same reference also reports an extensive simulation study aiming at comparing different methods within the framework, and the application to a brain imaging data set.

2 A unified framework for locally testing functional data

Assume to observe a sample of size n of continuous functional data defined on the domain $D \subset \mathbb{R}$. We refer to [2] for the generalization of this method to higher dimensional domains. Assume that we are interested in testing a null hypothesis H_0 against an alternative hypothesis H_1 in a local perspective. For each $t \in D$ we denote as H_0^t and H_1^t the restrictions of the null and alternative hypothesis at t . The unadjusted p -value function $p(t)$, $t \in D$, can be defined as in [7], by computing at each point t the p -value of the test of H_0^t against H_1^t . Our aim is to define an adjusted p -value function $\tilde{p}(t)$ controlling the FWER.

In the multivariate framework the FWER is defined as the probability of committing at least one type I error, i.e., the probability that at least one adjusted p -value associated to the components where the null hypothesis is true is below the chosen significance level $\alpha \in (0, 1)$. Similarly, in the functional framework, the FWER that can be defined as the probability that at least one point t of the domain such that H_0^t is true has an adjusted p -value lower than α . If $\forall \alpha \in (0, 1), \text{FWER} \leq \alpha$, the inferential procedure is provided with a strong control of the FWER. We achieve a weak control if the FWER is controlled assuming that H_0^t is true $\forall t \in D$.

The unified framework that we propose for adjusting the p -value function is summarized in Figure 1. It is a general procedure that requires choosing a family \mathcal{F} of subsets of the domain as a starting point. Some resulting procedures with different choices of \mathcal{F} will be detailed in the next section. The unified framework consists in the following steps:

1. Computation of p -values for all subsets. For all $S \in \mathcal{F}$, we compute the p -value p_n^S of the test of H_0^S against H_1^S . Such p -values can be computed using the integral over S of a point-wise test statistic, as discussed in [1].
2. Computation of the adjusted p -value function. For all $t \in D$, the adjusted p -value function is defined as the supremum of all p -values of sets including t .
3. Domain selection. The subsets of the domain where H_0 is rejected at level $\alpha \in (0, 1)$ are selected with a thresholding of the adjusted p -value function

The theoretical properties of the resulting procedure - in terms of FWER control and consistency - depend on the choice of the family \mathcal{F} , as shown in [2]. In particular, it is possible to show that if the family \mathcal{F} is fixed a priori, the procedure is in general provided with a control of the FWER which is intermediate between the weak and the strong control (unless the family is composed by all possible subsets of the domain). Furthermore it is possible to provide sufficient conditions on \mathcal{F} for providing an asymptotic strong control of the FWER.

3 Methods within the unified framework

Depending on the choice of \mathcal{F} , one can obtain very different procedures within the same unified framework. Here we briefly discuss some possible choices of the fam-

ily. We show how many functional testing procedures already well established in the literature of functional data analysis fall within the unified framework for some particular choices of the family \mathcal{F} . Finally, we briefly describe the threshold-wise testing, that is a novel testing procedure (see [2] for details) with sound theoretical properties.

Global testing. If we choose a family that only includes the whole domain, we obtain a global testing procedure. The global testing procedure is consistent and is provided with a weak control of the FWER. It is also the most powerful out of the ones presented. However, it does not provide domain selection, since the adjusted p -value is constant over the domain.

Borel-wise testing. The Borel-wise testing procedure is based on a family composed by all Borel sets of the domain D of non-null Lebesgue measure. The resulting procedure is the continuous extension of the closed testing procedure [6], that has been proposed in multivariate analysis to adjust p -values. Such procedure is provided with a strong control of the FWER. However, as shown in [2], this procedure is nearly useless in the practice, since its adjusted p -value function is also constant on D and is greater or equal to the maximum value of the unadjusted p -value function. Hence, such procedure is not consistent, and does not provide domain selection.

Partition closed testing. Assume that we have information about a partition of the domain into subsets of particular interest. Then the partition closed testing (first proposed in [9]) is the procedure based on the family containing all possible unions between sets of the partition. Such procedure is provided with a control of the FWER

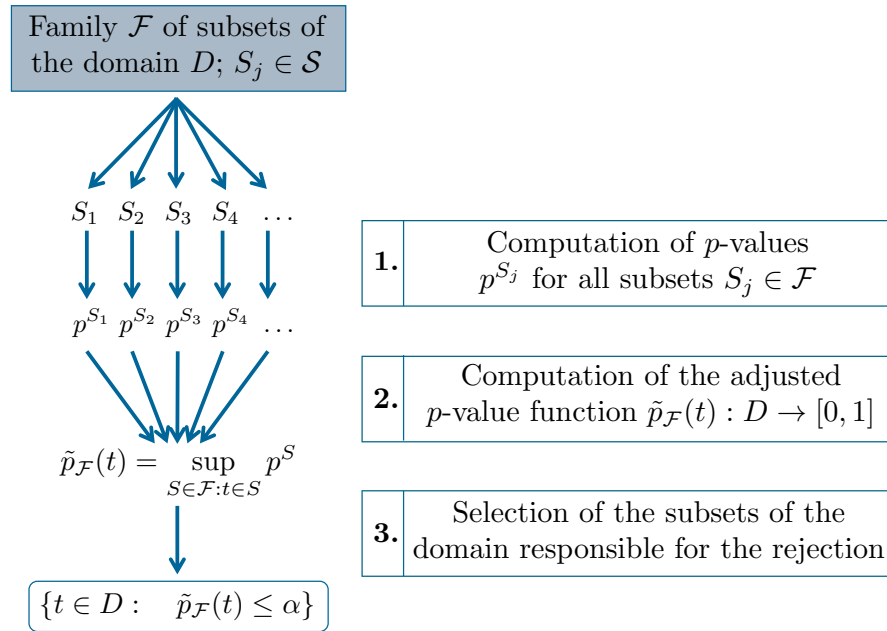


Fig. 1 Scheme of the unified framework for controlling the FWER on domain subsets

that is intermediate between the weak and the strong control: it is provided with a weak FWER control within subsets of the partition, and a strong FWER control between subsets of the partition.

Interval-wise testing. The interval-wise testing - first proposed in [7] - is based on a family composed by every interval of the domain. This method is also provided with an intermediate control of the FWER. In detail, in this case the control is strong within intervals but weak between intervals.

Threshold-wise testing. The idea behind the threshold-wise testing is to construct a family derived from the unadjusted p -value function, thus being data dependent. We define the family as made of the sublevel and superlevel sets of the unadjusted p -value function. This method is provided with a weak control of the FWER. Being based on a data driven family, it is not provided with a strong control of the FWER for finite sample sizes. However, as shown in [2], the procedure is provided with an asymptotic control of the FWER.

4 Conclusions

We conclude observing that in the case of pre-defined families it is not possible to guarantee both the possibility of performing domain selection and strong control of the FWER. Indeed, the only procedure able to control the FWER for finite samples is the Borel-wise testing, which, however, is nearly useless in the practice. This limitation introduces the urgency of focusing on data-driven families. Indeed, in this enlarged setting and for large sample sizes, it is possible to identify families that could both provide an (asymptotic) strong control of the FWER and allow for domain selection at the same time. As an example of this approach, a novel threshold-wise testing procedure is introduced in [2], and briefly described here.

References

1. Abramowicz, K., Häger, C.K., Pini, A., Schelin, L., Sjöstedt de Luna, S., Vantini, S.: Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scandinavian Journal of Statistics* **45**(4) (2018)
2. Abramowicz, K., Pini, A., Schelin, L., Sjöstedt de Luna, S., Stamm, A., Vantini, S.: Domain selection and family-wise error rate for functional data: a unified framework. Tech. Rep. 10/2019, MOX - Department of mathematics, Politecnico di Milano (2019)
3. Aneiros, G., Cao, R., Fraiman, R., Genest, C., Vieu, P.: Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis* **170**, 3–9 (2019)
4. Goia, A., Vieu, P.: An introduction to recent advances in high/infinite dimensional statistics. *Journal of Multivariate Analysis* **146**, 1–6 (2016)
5. Horváth, L., Kokoszka, P.: *Inference for functional data with applications*, vol. 200. Springer Science & Business Media (2012)
6. Marcus, R., Peritz, E., Gabriel, K.R.: On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**(3), 655–660 (1976)

7. Pini, A., Vantini, S.: Interval-wise testing for functional data. *Journal of Nonparametric Statistics* **29**(2), 407–424 (2017)
8. Ramsay, J.O., Silverman, B.W.: *Functional data analysis*. Springer, New York (2005)
9. Vsevolozhskaya, O.A., Greenwood, M.C., Powell, S.L., Zaykin, D.V.: Resampling-based multiple comparison procedure with application to point-wise testing with functional data. *Environmental and Ecological Statistics* **22**(1), 45–59 (2015)

Application and validation of dynamic Poisson models to measure credit contagion

Applicazione e validazione di modelli di Poisson dinamici per misurare il contagio nel credito

Arianna Agosto and Emanuela Raffinetti

Abstract The growing importance of financial technology platforms, based on interconnectedness, makes necessary the development of credit risk measurement models that properly take contagion into account. To this aim, we propose to use a credit risk model that allows to investigate contagion through Poisson autoregressive stochastic processes. We apply this model to the quarterly count of defaulted loans in the Italian banking system, finding evidence of contagion effects in several economic sectors. To calculate the accuracy of the model we use a new measure, whose main advantage is being not dependent on the type of predicted variable.

Abstract *La diffusione di piattaforme finanziarie digitali, basate sull'interconnessione, rende necessario lo sviluppo di modelli per il rischio di credito che tengano in opportuna considerazione il contagio. A tale scopo proponiamo di utilizzare un modello per il rischio di credito che permette di studiare il contagio attraverso processi di Poisson autoregressivi. Applicando il modello alle serie trimestrali del numero di prestiti a default nel sistema bancario italiano, troviamo evidenza di effetti di contagio in diversi settori economici. Per calcolare l'accuratezza del modello utilizziamo una nuova misura, il cui principale vantaggio risiede nella non dipendenza dalla tipologia di variabile risposta considerata.*

Key words: Credit Risk, Poisson Autoregressive models, Systemic risk, Accuracy measures

Arianna Agosto

Department of Economics and Management, University of Pavia, 27100 Pavia, Italy e-mail: arianna.agosto@unipv.it

Emanuela Raffinetti

Department of Economics, Management and Quantitative Methods, University of Milan, via Conservatorio 7, 20122 Milano, Italy e-mail: emanuela.raffinetti@unimi.it

1 Introduction

Interconnectedness between the economic agents was recognized as a trigger of the great financial crisis in 2008-2009 and is increasing with the growth of innovative financial technologies (fintechs), especially peer to peer lending platforms. Measuring the credit risk arising from interconnectedness is necessary to safeguard investors and maintain financial stability.

To this aim, we apply a novel credit risk model that incorporates contagion through Poisson autoregressive stochastic processes.

The main systemic risk measures defined in the literature (see e.g. Adrian and Brunnermeier, 2011 and Acharya et al., 2012) have been applied to financial market data and are based on Gaussian processes.

We indeed study contagion through discrete data models for default counts. Examples of default counts modelling can be found in Lando and Nielsen (2010), Koopman et al. (2012) and, recently, Azizpour et al. (2018). Our contribution is proposing a credit risk assessment approach based on a dynamic model which includes autoregressive, contagion and exogenous effects in a time-varying Poisson intensity specification.

Validation is a critical issue in credit risk modelling, because of the interest in selecting indicators able to predict the default peaks. In our empirical application we validate the models applied to default counts using a newly developed predictive accuracy measure. Contrary to the main summary predictive accuracy indices, such as the root mean squared error and the area under the ROC curve, the new measure, which is called Rank Graduation measure, does not depend on the type of response variable being predicted. It is based on the distance between the observed response variable values and the same values re-ordered in terms of the predicted values given by the model and was found quite effective in two real machine learning applications characterised, respectively, by a binary and a continuous response variable. Our aim is to show how the Rank Graduation measure appears as an appropriate model predictive accuracy index also when dealing with discrete response variables, here represented by count variables.

The paper is organised as follows. Section 2 describes the proposed modelling approach. Section 3 provides the basic elements characterising the Rank Graduation measure. Section 4 presents the main empirical findings derived from the application and validation of the presented model for default counts. Section 5 concludes.

2 Proposal

We propose to use Poisson autoregressive models to investigate inter-sectoral correlation. Specifically, Agosto and Giudici (2019) have defined the following model for the number of defaults in economic sector j at time t :

$$\begin{aligned}
y_{jt} | \mathcal{F}_{t-1} &\sim \text{Poisson}(\lambda_{jt}) \\
\log(\lambda_{jt}) &= \omega_j + \sum_{i=1}^p \alpha_{ji} \log(1 + y_{jt-i}) + \sum_{i=1}^q \beta_{ji} \log(\lambda_{jt-i}) \\
&\quad + \sum_{i=1}^r \gamma_{ji} x_{t-i} + \sum_{i=1}^s \zeta_{ji} \log(1 + y_{kt-i})
\end{aligned} \tag{1}$$

with $\omega_j, \alpha_{ji}, \beta_{ji}, \gamma_{ji}, \zeta_{ji} \in \mathbb{R}$ and $x_{t-i} := (x_{1t-i}, x_{2t-i}, \dots, x_{dt-i})' \in \mathbb{R}^d$ being a vector of lagged exogenous covariates. In other words it is assumed that the default count time series y_{jt} in a given sector called j , conditional on its past, follows a Poisson distribution with a time-varying autoregressive intensity λ_{jt} whose formulation also includes the past default counts y_{kt-i} of another sector called k and, possibly, a set of exogenous covariates x_t . The impact of other sectors' default counts is what we call contagion. Taking the $\log(\cdot) + 1$ of the counts makes possible to deal with zero values. The inclusion of past values of intensity λ_{jt} allows for parsimonious modelling of long memory effects in a way analogous to the extension of standard ARCH processes to GARCH ones for Gaussian variables. It can be shown that including an autoregressive component in a Poisson process generates overdispersion (i.e. unconditional variance larger than the mean), an empirical feature of default count series, which typically cluster in time. The exogenous component x_t contains macroeconomic and financial variables affecting all companies' default probability, i.e. *systematic risk factors*.

Model (1) is called Contagion PARX as it is inspired by Poisson Autoregression with Exogenous Covariates (PARX) developed by Agosto et al. (2016), who studied the properties of the process and applied it to Moody's rated US default counts. Differently from the PARX and following Fokianos and Thj\o stheim (2011), Contagion PARX uses a log-linear specification, rather than a linear one, for the intensity, allowing to consider negative dependence on the covariates, a likely possibility in application to default risk. Furthermore, the model includes an explicit contagion component, i.e. the default count in other sectors, while Agosto et al. (2016) only considered an autoregressive part and an exogenous (macroeconomic) one.

3 A brief overview of the Rank Graduation index

With the aim of assessing the predictive accuracy associated with the models described in the application section, an overview on the *RG* index is here presented. When the response variable Y is both binary and continuous there is not a unique measure. Giudici and Raffinetti (2019) have worked out one possible solution. The proposal is based on the C concordance curve, obtained by ordering the Y original values according to the ranks of the corresponding \hat{Y} estimated values. Let Y be a (binary, continuous or also, as in the case discussed in this paper, discrete) response variable and let X_1, \dots, X_p be a set of p explanatory variables. Suppose to apply a model such that $\hat{y} = f^h(\mathbf{X})$, with $h = 1, \dots, l$. The model predictive accuracy is assessed by measuring the distance between the set of points lying on the C concordance curve $(i/n, (1/(n\bar{y})) \sum_{j=1}^i y_{\hat{r}_j})$, where $\bar{y} = \sum_{i=1}^n y_i$ and $y_{\hat{r}_j}$ represents the j -th

response variable value ordered by the rank of the corresponding predicted value, and the set of points lying on the bisector curve $(i/n, i/n)$ (see Figure 1).

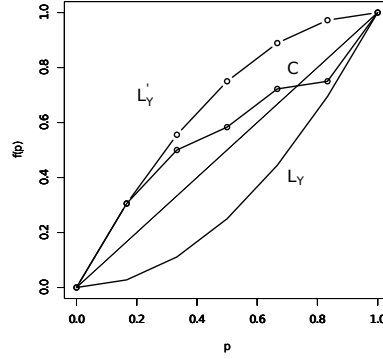


Fig. 1 The L_Y and L_Y' Lorenz curves and the C concordance curve, normalised.

If the C concordance curve overlaps with the bisector curve, the model has no predictive capability. On the contrary, if the C concordance curve overlaps with the response variable Lorenz curve L_Y (defined by the normalised Y values ordered in non-decreasing sense) or the dual Lorenz curve L_Y' (defined by the normalised Y values ordered in non-increasing sense), situations of concordance and discordance between Y and \hat{Y} arise, meaning that the model perfectly explains the target variable.

The RG (Rank Graduation) index is defined as:

$$RG = \sum_{i=1}^n \frac{\left\{ (1/(n\bar{y})) \sum_{j=1}^i y_{\hat{r}_j} - i/n \right\}^2}{i/n} = \sum_{i=1}^n \frac{\{C(y_{\hat{r}_i}) - i/n\}^2}{i/n}, \quad (2)$$

where $C(y_{\hat{r}_j}) = \frac{\sum_{j=1}^i y_{\hat{r}_j}}{\sum_{i=1}^n y_{r_i}}$ represents the cumulative values of the (normalised) response variable.

Note that the RG index takes values between 0 and RG_{max} , which is obtained when the predicted ranks order the response variable values in full concordance (or full discordance) with the observed ranks. In addition, we remark that when some of the \hat{Y} values are equal to each other, the original Y values associated with the equal \hat{Y} values are substituted by their mean, as suggested by Ferrari and Raffinetti (2015).

4 Application

In this section we provide the application of the model presented in Section 2 to Italian corporate default counts data.

As a proxy of default counts, we use the number of transitions to *bad loans* of Italian banks' credit exposures collected in Bank of Italy's Credit Register. Bad loans are exposures to insolvent debtors that the bank does not hope to recover anymore and must report as losses in its balance sheet. The data are quarterly and divided by economic sector. We use data covering the period March 1996 - June 2018 (90 observations).

Figure 2 shows the default count time series of two economic sectors of major importance: Households and Commercial corporate sectors. Both series exhibit clustering and a possible structural break in 2009, with an increase in both level and variability.

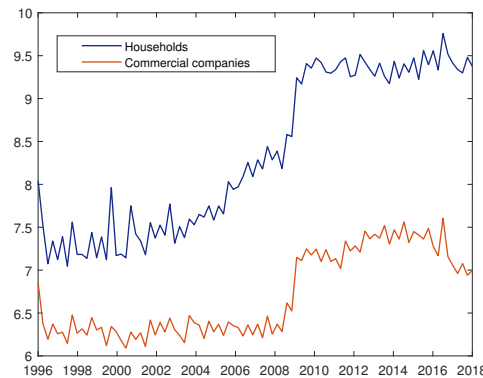


Fig. 2 Default count time series in the Italian banking system (logarithmic scale)

To investigate credit contagion effects, we apply model (1) to various economic sector default counts. As the number of available observations is limited, we carry out a sector pairwise analysis. Specifically, in each regression the dependent variable is the number of defaults in a given sector (y_{jt} in specification 1), while the covariate is the number of defaults in another sector (y_{kt}), lagged by one or two time periods. That is for now we use model (1) with $\gamma_{ji} = 0$. The results are shown in Table 1. Each row represents a model, in which the default counts of one sector are pairwise regressed on all the others, lagged by one or two periods, and on the autoregressive component, according to the presented specification.

From Table 1 note that Commercial and Financial sectors are the most influenced by the default dynamics of the others. As expected from an economic point of view, Households, which represent the consumer sector, influence all the other sectors. An interesting interaction is that involving Real Estate and Households sectors: the first, including both constructions and real estate companies, is affected by Households default risk, as expected, but in turn affects Households, which are typically highly engaged in the housing market.

	HH	MNF	RE	FIN	CMM
HH			0.093		
MNF	0.086				
RE	0.203	0.252		0.650	
FIN	0.298	0.275	0.925		0.926
CMM	0.267	0.264	0.473		

Table 1 Significant effects between sectors, from the fitted Contagion PARX models (HH=Households, RE=Constructions and Real Estate, FIN=Financial, CMM=Commercial). The rows report the infected sectors (as response variables) and the columns the infecting sectors (as explanatory lagged variables). The reported effects are the sum of the first and second lag coefficients).

To provide an example of validation, we consider the model regressing Commercial sector default counts on their past values and on Household default counts, which is of major economic interest.

We then compare three different specifications. The first - that we call Full Model - includes both the contagion component and a macroeconomic covariate. The covariate we consider is the Italian GDP quarterly growth rate, which has turned out to be the preferable exogenous regressor through a preliminary model selection. In other words, referring to formulation (1), GDP growth rate is the x process while the Households default counts are the y_k one. The addition of an exogenous covariate in the intensity specification can be considered as a robustness step of our contagion analysis, verifying to what extent the macroeconomic and financial stress affecting all economic agents explains the default and contagion dynamics. The second model is the Contagion Model without other covariates than Households default counts (γ parameters equal to 0 in specification 1, like the models in Table 1). The third (PAR Model) is only autoregressive, without covariates (both γ and ζ parameters equal to 0 in formulation 1).

We compare the in-sample performances of the three models above by using the RG measure. The RG index computed on Full Model equals 3.826 against a value of 3.627 for PAR Model. Contagion Model provides a value of 3.790 for the RG index. It thus follows that PAR Model explains the 83% of the variable ordering, compared with the 86% of Contagion Model and the 87% of Full Model. This means that adding the contagion component leads to a remarkable increase in the predictive accuracy with respect to the only autoregressive specification. GDP turns out to be a significant regressor for default intensity: the coefficient associated to the second lag of GDP growth rate is -0.081 with a t-statistic of -3.340. Though, its inclusion improves model performance only slightly.

5 Conclusion

We have proposed a credit risk modelling approach which allows to investigate contagion through Poisson autoregressive stochastic processes.

Applying the contagion model to the quarterly count of defaulted loans in the Italian banking system, we find evidence of significant inter-sectoral contagion effects. The model is also found to have a high predictive accuracy. Considering the macroeconomic context (through the GDP growth rate) improves the model performance, even not drastically. This suggests that both contagion effects and exogenous factors should be taken into account for default prediction.

Our approach may represent an attractive tool in the current smart world, where the widespread development of innovative financial technologies (fintechs) makes available big amounts of data on financial behaviours and interconnections.

The availability of large datasets would also support future research on application and validation of multivariate multi-response contagion models.

References

1. Acharya, V., Engle, R., Richardson, M.: Capital Shortfall: A New Approach to Ranking and Regulating Systemic Risks. *American Economic Review: Papers and Proceedings* **102**, 59–64 (2012)
2. Azizpour, S., Giesecke, K., Schwenkler, G.: Exploring the sourcing of default clustering. *Journal of Financial Economics* **129** (1), 154–183 (2018)
3. Adrian, T., Brunnermeier, M.K.: CoVaR. *American Economic Review: Papers and Proceedings* **106**, 1705–1741 (2016)
4. Agosto, A., Cavaliere, G., Kristensen, D., Rahbek, A.: Modeling Corporate Defaults: Poisson Autoregressions with Exogenous Covariates (PARX). *Journal of Empirical Finance* **38**(B), 640–663 (2016)
5. Agosto, A., Giudici, P.: Dynamic Poisson models to measure credit default contagion. Submitted (2019)
6. Ferrari, P.A., Raffinetti, E.: A different Approach to Dependence Analysis. *Multivar. Behav. Res.* **50**, 248–264 (2015)
7. Fokianos, K., Tjøstheim, D. : Log-linear Poisson autoregression. *Journal of Multivariate Analysis* **102**, 563–578 (2011)
8. Giudici, P., Raffinetti, E.: A Rank Graduation measure to assess predictive accuracy. To be submitted (2019)
9. Koopman, S.J., Lucas, A., Schwaab, B., 2012: Dynamic factor models with macro, frailty, and industry effects for U.S. default counts: the credit crisis of 2008. *Journal of Business and Economic Statistics* **30**, 521–532 (2012)
10. Lando, D., and Nielsen, M.: Correlation in corporate defaults: Contagion or conditional independence? *Journal of Financial Intermediation* **19**, 355–372 (2010)

Monitoring SDGs at territorial level: the case of Lombardy

Il monitoraggio degli SDGs a livello territoriale: il caso della Lombardia

Leonardo Alaimo, Livia Celardo, Filomena Maggino, Adolfo Morrone, Federico Olivieri

Abstract: In this paper we want to measure the trend of Lombardy compared to the Italian average in each goal of sustainable development. By doing this, we propose to synthesize a large set of sustainable development indicators at Italian regional level. Complexity represents the biggest challenge in monitoring the 2030 Agenda. In this perspective, composite indicators represent a useful tool that allows a quick and concise view of performances related to each goal. The intention is to provide stakeholders and media with synthetic, clear and easy-to-read evaluations of performances of each region to provide an insight on the direction the regions are heading to and if they are going in the right direction towards the achievement of the SDGs.

Abstract: *In questo lavoro proponiamo la sintesi di un vasto set di indicatori di sviluppo sostenibile a livello regionale al fine di misurare per ogni obiettivo il trend della Lombardia rispetto alla media Italiana. La complessità rappresenta la più grande sfida nel monitoraggio dell'agenda 2030. In questa prospettiva, gli indici composti rappresentano un utile strumento in grado di mostrare in maniera concisa le performances di ogni goal. L'idea è quella fornire agli stakeholders una valutazione sintetica e chiara delle prestazioni di ogni regione, in termini di raggiungimento degli obiettivi (SDGs).*

Leonardo Alaimo
e-mail: leonardo.alaimo@uniroma1.it
Livia Celardo
e-mail: livia.celardo@uniroma1.it
Adolfo Morrone
e-mail: adolfo.morrone@aics.gov.it
Federico Olivieri
e-mail: federico.olivieri@asvis.net

Key words: Sustainable development; SDGs; Lombardy; Composite Indicators; AMPI

1 Introduction

In September 2015, 193 countries adopted the 2030 Agenda for Sustainable Development and its 17 Sustainable Development Goals (hereinafter: SDGs). The SDGs are the blueprint to achieve a better and more sustainable future for all. They address the global challenges we face, including those related to poverty, inequality, climate, environmental degradation, prosperity, peace and justice.

The Agenda recognizes the importance of territories in the implementation of policies required to reach the SDGs. In Italy the national strategy for sustainable development underline the key role of regions in the implementation of the Agenda 2030¹. Many studies analysed and monitored the differences among Italian Regions in achieving SDGs (Authors, 2018, Authors, 2018, Author, 2019).

In order to support this process, the Italian Alliance for the Sustainable Development (ASviS) is committed to monitor the achievement of the SDGs at regional level. Starting from the UN indicators statistical framework, ASviS selected through a structured dialogue with 217 partners from the civil society a set of statistical indicators to measure the SDGs at regional level. The result of this work is the 2018 ASviS report, in which the monitoring of SDGs at regional level was addressed for the first time.

In this paper, we analyse the positioning of Lombardy related to the SDGs, using as frame of reference the Italian trend. We use composite indicators as method to represent the level of sustainable development in the Region.

2 Data and methods

We used 77 basic indicators, divided among the different goals, all in time series from 2010 to 2016. The source of the data is the ASviS dataset², which includes only indicators from official statistical sources (Istat, Ispra, etc.). This dataset is the result of an intense dialogue with stakeholders and it is a tool for monitoring the positioning of regions compared to Italian average. Nevertheless, in order to calculate composite indicators, we had to perform a selection of the basic indicators

¹ <https://www.minambiente.it/pagina/la-snsvs>

² <http://asvis.it/il-monitoraggio-degli-sdgs-a-livello-regionale/>

Monitoring SDGs at territorial level

that was influenced by the need to have data in time series and available at regional territorial disaggregation level.

As previously written, the main purpose of this article is analysing the situation of Lombardy with respect to each goal using one single composite measure and compare it with the Italian situation³. To do this, we summarized each set of basic indicators in composite indicators (hereinafter: composites) through a method that will be illustrated below. From the operational point of view, after the definition of the phenomenon and the selection of basic indicators, the following phases are the normalization of the individual indicators and the aggregation of the normalized indicators (Mazziotta and Pareto, 2016).

For the aggregation, we chose the Adjusted Mazziotta-Pareto Index (AMPI), a partially non-compensatory composite indicator also used by Istat for the construction of equitable and sustainable well-being (BES) indicators and previously used by ASviS for measuring the trends of each goal at Italian and European level (ASviS, 2018). It is a variant of Mazziotta Pareto Index (MPI), based on a Min-Max normalisation and a re-scaling of the basic indicators in a range [70; 130], according to two goalposts, representing a minimum and a maximum value of each variable for all units and time periods (Mazziotta and Pareto, 2016). Using this normalisation procedure of the individual indicators allows assessing absolute changes over time. AMPI allows computing the score of each unit independently of the others, in contrast to the MPI where the mean and standard deviation of the individual indicators are required (Mazziotta and Pareto, 2017, 179). Given the original matrix (1):

$$X = \{x_{ij}\} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix} \quad (1)$$

where $i=1,...,n$ are the units of analysis and $j=1,...,m$ are the variables, we calculate the normalized matrix as follows (2):

$$r_{ij} = \frac{(x_{ij} - \text{Min}_{x_j})}{(\text{Max}_{x_j} - \text{Min}_{x_j})} * 60 + 70 \quad (2)$$

where x_{ij} is the value of the indicator j in the unit i and Min_{x_j} and Max_{x_j} are the goalposts for the indicator j ⁴. In the normalization, it is necessary to define the polarity of the basic indicators, i.e. the sign of the relation between the indicator itself and the phenomenon to be measured. Therefore, the type of composite we

³ For the calculation of composite indices, we have taken into account all Italian regions, not only Italy and Lombardy.

⁴ Let Inf_{x_j} and Sup_{x_j} be the minimum and the maximum of indicator j across all time periods considered, and Ref_{x_j} be the reference value for indicator j . Then the “goalposts” are defined as: $\text{Ref}_{x_j} \pm \Delta$, where $\Delta = (\text{Sup}_{x_j} - \text{Inf}_{x_j})/2$ (Mazziotta and Pareto, 2017:178).

want to construct defines polarity. If the basic indicator has positive polarity, the formula (2) is used; if it has negative polarity, we calculate the complement to 200 of the (2). The polarity of basic indicators used in this paper is reported in the ASviS Report (ASviS, 2018).

In this article, the goalposts have been constructed using as reference the value assumed by Italy in the year 2010 for each basic indicator considered. Thus, using this normalization procedure, each indicator will assume the value 100 for Italy in 2010 and all the other values, of each unit for all the years, will be expressed in reference to this value, allowing a comparison in time and space. The AMPI is given by (3):

$$AMPI^{\pm} = \mu_{r_i} \pm \sigma_{r_i} * cv_i \quad (3)$$

where μ_{r_i} , σ_{r_i} and $cv_i = \sigma_{r_i} / \mu_{r_i}$ are the mean, the standard deviation and the coefficient of variation of the unit i and the sign \pm depends on the kind of phenomenon measured. In this paper, all the composites are positive, i.e., increasing values of each index correspond to positive variations of the phenomenon considered in each goal; then we used AMPI with negative penalty (AMPI-). All values will be approximately within the range [70,130], and 100 will represent the reference value (in our case, that of Italy in 2010). Therefore, AMPI indicates how each unit is placed with respect to the goalposts.

3 Results

Figure 1 reports the charts with the time series of the composites for each goal considered; the value of Lombardy is compared to the national data.

Monitoring SDGs at territorial level

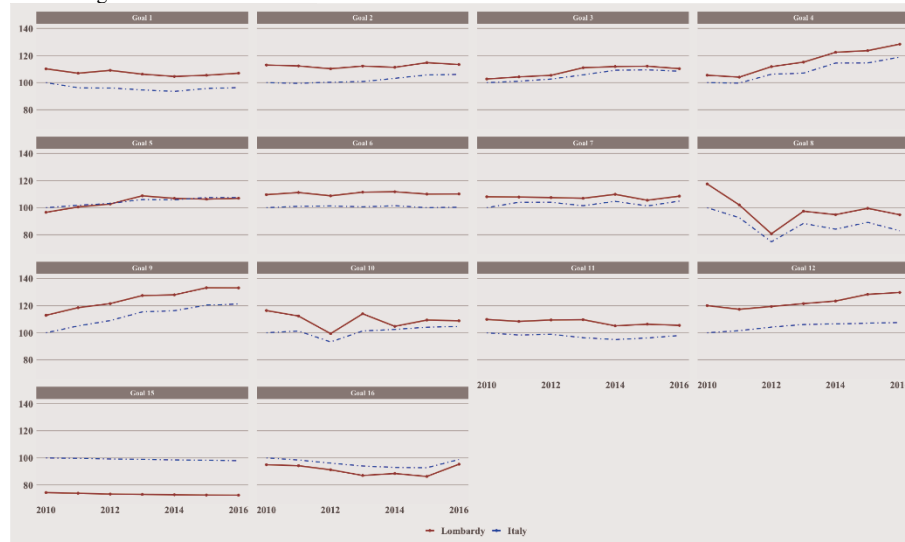


Figure 1: SDGs composite indicators: data of Lombardy and Italy. time series 2010–2016; AMPI: Italy 2010 = 100.

The composites for goals 1 (Poverty), 2 (Hunger), 4 (Education), 6 (Clean water and sanitation), 7 (Affordable and clean energy), 8 (Decent work and economic growth), 9 (Industry, innovation and infrastructure), 10 (Reduce inequalities), 11 (Sustainable cities and communities), 12 (Responsible production and consumption) reveal for Lombardy a better situation than the Italian average. The positive performance in goal 1 is attributable to the superior situation of its basic indicators: for instance, severely materially deprived people, in 2016 reaches 6% compared to 12% of the country, poor households is equal to 5% in comparison to 11% of Italian average. The increasing trend of education (SDG 4) can be explained by the broad rise of the tertiary educational attainment that increases from 23% in 2010 to 34% in 2016 (Italy reaches 27% in 2016). The higher level of SDG 6 is explained by the degree of the efficiency of urban water supply network, 73% compared to the 59% of national average. The composite for SDG 12 performs better than the Italian average due to the higher level of separate collection of municipal waste (68% in Lombardy compared to the 52,5% of the national level). The composite indicator of SDG 15 (Life on land) shows a worse situation compared to the Italian average. This is explained by the levels of the land fragmentation, 52,4%, in comparison to the 38% of Italy, and by those of soil cover, 13% compared to the 8% of national average. The lower level of SDG 16 is mainly due to the higher number of thefts, in Lombardy 258 in comparison to the 222 of Italy. The composites for gender equality (SDG 5) and good health and well-being (SDG 3) are in line with the national data. It wasn't possible to elaborate a composite indicator for SGDS 13 (climate action), 14 (life below water) and 17 (partner for the Goals), due to the lack of data at regional level.

4 Conclusions

Sustainable development, apart from being a central theme in the international debate, is today an essential necessity that must guide the definition of policies and the choices of allocation and use of resources. From this point of view, a central role is played by the territorial realities, in order to be able to define and calibrate policies and actions in the best possible way. Italy has always been characterized by deep differences between the various areas of the country, which should lead to specific interventions for specific problems. In this paper, we have taken into consideration the case of Lombardy. The analysis identified the differences existing between the Region and the national data, making explicit some goals with better and other with worse values than Italian ones.

References

1. Authors. (2018). "Sviluppo sostenibile e differenze regionali." AIQUAV 2018. V Convegno dell'Associazione Italiana per gli Studi sulla Qualità della Vita. Libro dei Contributi Brevi, pp. 199-206. Genova: Genova University Press.
2. Author. (2019). "Sustainable Development and Territorial Differences: an Italian Analysis of Economic Sustainability."
3. ASviS. (2018). L'Italia e gli Obiettivi di Sviluppo Sostenibile. Rapporto ASviS 2018.
4. Authors. (2018). "Gli indicatori regionali per lo studio delle disuguaglianze economiche." *Energia, Ambiente e Innovazione*, Vol. 3/2018, pp. 128–135. Doi: 10.12910/EAI2018-070.
5. Mazziotta, M., Pareto A. (2016). "On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phenomena." *Soc. Indic. Res.*, 127(3), 983–1003.

The Experts Method for the prediction of periodic multivariate time series of high dimension

Il Metodo degli Esperti per la previsione di serie temporali multivariate e periodiche, di dimensione elevata

Giacomo Aletti, Marco Bellan and Alessandra Micheletti

Abstract We propose a method, called *Experts Method*, to predict the evolution of a high dimensional multivariate set of time series. The method is based on the definition of a set of "experts", which are portions of a training set of the considered time series which best fit the data immediately preceding those to be predicted. A suitable combination of Singular Value Decompositions is used to filter out the noise, and provide robust predictions. The advantage of this method, if compared with classical multivariate time series analysis, is that it can be applied also when the time series column order is reshuffled, from time to time, in the collected dataset.

Abstract *In questo lavoro viene proposto un metodo, detto Metodo degli Esperti, per predire l'evoluzione di un insieme multivariato e di grosse dimensioni di serie temporali. Il metodo é basato sulla definizione di un insieme di "esperti", ovvero di porzioni, di un training set delle serie temporali considerate, che approssimano al meglio i dati che precedono quelli che devono essere previsti. Viene utilizzata una opportuna combinazione di decomposizioni ai valori singolari per filtrare il rumore, e fornire previsioni robuste. Il vantaggio di questo metodo, rispetto ai classici metodi di analisi di serie temporali multivariate, é che esso può essere applicato anche quando l'ordine con cui le serie sono registrate nelle colonne del dataset, viene scambiato di tanto in tanto.*

Key words: SVD, prediction, multivariate time series

Giacomo Aletti, Alessandra Micheletti

Dept. of Environmental Science and Policy, Università degli studi di Milano e-mail: giacomo.aletti@unimi.it, alessandra.micheletti@unimi.it

Marco Bellan

Dept. of Mathematics, Università degli studi di Milano e-mail: marco.bellan@studenti.unimi.it

1 Introduction

In this paper we describe a method, called *Experts Method*, to forecast the evolution of a multivariate set of time series, of big dimension, and with partially censored data. It has been applied to the data provided by the H2020 Big Data Horizon Prize 2017 [1]. The data subject to the forecast are energy flow-related measurements over $N_T = 1912$ high-tension lines registered over two years (called *target data*). The lines have been anonymized with respect to location, thus they are provided as purely temporal data. The energy flows are registered every 5 minutes for one year, but since the data recording can be switched off for some (random) period of time, some of the time series show missing data. The data are divided into subsequent files, where each column of the files contains the observations of one high-tension line in the time period $[t, t + \Delta t]$. The length Δt of the periods of observation can be different from file to file (it anyway is always bigger or equal than 1 hour). This setting emulates the fact that in real applications data are recorded continuously by sensors, but they can be passed to the statistical analysis during different times, either regularly or when the measurements overcome a fixed threshold, or some extra control is planned and an immediate forecast of the future behaviour is needed.

The Experts Method here proposed is thus suitable to predict the evolution of a multivariate set of *data streams* of big dimension, where the observed streams are given during different random times.

The method here proposed is supervised, thus the target data have been divided into a training set, to set up the method, and an adapt set, on which the method is tested. Anyway, differently from deep learning methods based on neural networks, our method does not need a huge training set to be trained. Moreover our method allows to predict the behaviour of the time series even if their order in the columns of the provided files is reshuffled from time to time.

The presence of missing data in the target has been taken into account as follows:

- in the training set the NaN are substituted with zeros, the +inf values (if any) are substituted with the maximum floating point value in double precision, the -inf values (if any) are substituted with the minimum floating point value in double precision.
- In the adapt set if the NaN's sequence is in between two observed values of the column in the file, we interpolate linearly the missing data between the two observed values. If the NaN's sequence is starting at the beginning of the file, we fill all the missing values with the first observed value in the column. If the NaN sequence is located at the end of the file, we fill the missing values with the last observed value.

In the following we describe the theoretical foundation of the Expert Method, and the implementation scheme.

2 Weigthed Sobolev discrete space

Given two bounded, piece-wise differentiable functions f and g on a fixed interval $[t_0, t_1]$, we will measure the distance between them through the weighted Sobolev distance d_S so defined:

$$d_S^2(f, g) := \int_{t_0}^{t_1} (f(t) - g(t))^2 dt + w \int_{t_0}^{t_1} (f'(t) - g'(t))^2 dt. \quad (1)$$

We will often refer to sets of equally sampled functions $\{f_j(t_i), 0 \leq i < M, 0 \leq j < N\}$, which form matrices $F = [f_j(t_i)]_{i,j}$. Thus, in our application, M is the number of considered time steps, N is the number of observed high-tension lines and $f_j(t_i)$ is the energy flow measured on line j at time t_i . The matrix $\Delta F = [f_j(t_i) - f_j(t_{i-1})]_{i,j}$ will play the role of estimating the piecewise constant derivative of each function and will be appended to F to form the extended matrix $F^{(\text{ext})}$:

$$F^{(\text{ext})} = \begin{pmatrix} F \\ \alpha \Delta F \end{pmatrix}, \quad \text{where } \alpha^2 = w(1 - \frac{1}{M}).$$

With this notation, given two matrices $F^{(1)}$ and $F^{(2)}$ of the same form, the Froebenious norm of the difference between the two extended matrices relates to the sum of the Sobolev squared distance of the corresponding functions:

$$\begin{aligned} & \|F^{(1)(\text{ext})} - F^{(2)(\text{ext})}\| \\ &= \sum_j \left(\sum_{i < M} (f_j^{(1)}(t_i) - f_j^{(2)}(t_i))^2 \right. \\ & \quad \left. + \alpha^2 \sum_{0 < i < M} ((f_j^{(1)}(t_i) - f_j^{(1)}(t_{i-1})) - (f_j^{(2)}(t_i) - f_j^{(2)}(t_{i-1})))^2 \right) \\ & \simeq \sum_j \left(\int_{t_0}^{t_1} (f_j^{(1)}(t) - f_j^{(2)}(t))^2 dt + w \int_{t_0}^{t_1} (f_j^{(1)'}(t) - f_j^{(2)'}(t))^2 dt \right) \\ &= \sum_j d_S^2(f_j^{(1)}, f_j^{(2)}). \end{aligned}$$

We will use the Sobolev squared distance to identify the subset of the training set which fits better with the last observations of the electric lines before the prediction.

3 Implementation of the Experts Method

We assume to receive in input a matrix B of dimension $M_A \times N_A$

$$B = \{B_{t_i,j}\}, \quad 0 \leq i < M_A, 0 \leq j < N_A,$$

where again M_A is the number of recorded time steps, and N_A is the (total) number of electric lines. We want to predict $\{B_{t_i,j}, M_A \leq i < M_A + \text{Hor}, 0 \leq j < N_A\}$, that is we want to predict the lines energy flow in a number Hor of time steps forward.

We extract the days of the week, the hours and the minutes of time spanned by B , and we extract from the training set the family of all the observations, called “experts”, $E = \{E^{(k)}\}$ that insist on the same timespan, but in another year and month,

$$E^{(k)} = \{E_{t_i,j}^{(k)}\}, \quad 0 \leq i < M_A + \text{Hor}, 0 \leq j^* < N_T, 0 \leq k < N_{\text{exp}}.$$

In this way we form a sample of “experts”, based on the observation that our data show a periodicity, with a period of one week. This is reasonable, since the variation in the request of electric energy is mainly related with the working habits of a region and their weekly recurrence. We neglect here the effects of yearly seasonality (differences from summer and winter), because we have a training set composed only by one year of observations. Also this type of seasonality could be taken into account in presence of richer datasets.

We do not assume that $N_T = N_A$, since the recording of some lines can have been switched off and we also assume that the columns of B can be reshuffled in subsequent files. The Experts Method guarantees a good prediction performance also when the recorded lines are stored into different columns of the data matrix at different time frames.

3.1 Approximation by experts

Given an expert $E^{(k)}$, we use

1. the first part $E^{(k,1)} = \{E_{t_i,j}^{(k)}, 0 \leq i < M_A, 0 \leq j < N_T\}$ to fit $\{B_{t_i,j}, 0 \leq i < M_A, 0 \leq j < N_T\}$.
2. the second part $E^{(k,2)} = \{E_{t_i,j}^{(k)}, M_A \leq i < M_A + \text{Hor}, 0 \leq j < N_T\}$ to predict the unknown $\{B_{t_i,j}, M_A \leq i < M_A + \text{Hor}, 0 \leq j < N_T\}$.

We compute the extended matrices of experts $E^{(k,1)(\text{ext})}$, and then, in order to fit each line $\{B_{t_i,j}, 0 \leq i < M_A\}$ (without overfitting it), we take the best rank $r < \min(N_T, M_A)$ approximation of each extended expert by computing the SVD decomposition [2, 3] of each $E^{(k,1)(\text{ext})}$. Accordingly, given

$$E^{(k,1)(\text{ext})} = \sum_{v=1}^{\min(N_T, M_A)} \mathbf{u}_v^{(k)} s_v^{(k)} \mathbf{v}_v^{(k)\top}$$

The Experts Method

with $s_1^{(k)} \geq s_2^{(k)} \geq \dots \geq s_{\min(N_T, M_A)}^{(k)} \geq 0$, we first remove the unnecessary noise (where $\frac{s_v^{(k)}}{s_1^{(k)}} < \text{factor}$). Then we keep only the first r components of the SVD decomposition, defining the smoothed experts

$$S_E^{(k,1)(\text{ext})} = \sum_{v=1}^r \mathbf{u}_v^{(k)} s_v^{(k)} \mathbf{v}_v^{(k)\top}.$$

We note that $S_E^{(k,1)}$ is the best rank r approximation, using the Sobolev distance, of the functions on the columns of each $E^{(k,1)}$.

3.2 Prediction from each expert

We predict each column of B independently of the others, thus we fix the column $\tilde{\mathbf{B}}_j = \{B_{ti,j}, 0 \leq i < M_A\}$ and we compute the extended column $\tilde{\mathbf{B}}_j^{(\text{ext})} = \{B_{ti,j}^{(\text{ext})}, 0 \leq i < M_A\}$. Each smoothed expert is requested to fit linearly $\tilde{\mathbf{B}}_j$ with its best RED-rank linear combination, where $\text{RED} \leq r$ is a further parameter chosen to reduce the dimensionality of the problem. Accordingly, given $X_{v,j}^{(k)} = \mathbf{u}_v^{(k)} \cdot \tilde{\mathbf{B}}_j^{(\text{ext})}$, each $X_{v,j}^{(k)}$ represents the projection of $\tilde{\mathbf{B}}_j^{(\text{ext})}$ on the orthonormalized vectors $\mathbf{u}_v^{(k)}$. Then, for any k, j , we take the RED indices $\{v_{j,1}^{(k)}, \dots, v_{j,\text{RED}}^{(k)}\} \subseteq \{1, \dots, r\}$ which maximize $(X_{v,j}^{(k)})^2$, and we solve the least squares problem

$$\mathbf{x}_j^{(k)} = \arg \min_{\mathbf{x} \in \mathbb{R}^{N_T}} \left\| \underbrace{\left(\sum_{l=1}^{\text{RED}} \mathbf{u}_{v_{j,l}^{(k)}}^{(k)} s_{v_{j,l}^{(k)}}^{(k)} \mathbf{v}_{v_{j,l}^{(k)}}^{(k)\top} \right) \cdot \mathbf{x} - \tilde{\mathbf{B}}_j^{(\text{ext})}}_{\text{RED approx. of } S_E^{(k,1)(\text{ext})}} \right\|^2.$$

The solution has the analytic expression $\mathbf{x}_j^{(k)} = \sum_{l=1}^{\text{RED}} \frac{X_{v_{j,l}^{(k)},j}^{(k)}}{s_{v_{j,l}^{(k)}}^{(k)}} \mathbf{v}_{v_{j,l}^{(k)}}^{(k)}$ for each expert k . A first prediction of $\{B_{ti,j}, M_A \leq i < M_A + \text{Hor}\}$ is hence based on $E^{(k,2)}$ and it is given by

$$\check{B}_{ti,j}^{(k)} = \sum_{j^*} E_{ti,j^*}^{(k)} x_{j^*,j}^{(k)}, \quad M_A \leq i \leq M_A + \text{Hor}.$$

Finally, to ensure that the prediction will start from the last observed value in the adapt file, a correction term is added $\hat{B}_{ti,j}^{(k)} = \check{B}_{ti,j}^{(k)} + (B_{t_{M_A}-1,j}^{(k)} - \sum_{j^*} E_{t_{M_A}-1,j^*}^{(k)} x_{j^*,j}^{(k)}) * h(t_i - t_{M_A} - 1)$, where h is a decreasing non-negative function with $h(0) = 1$.

3.3 Final prediction

For each column j of B , each expert k has given its prediction $\{\hat{B}_{t_i,j}^{(k)}, M_A \leq i < M_A + \text{Hor}\}$. Among them, we extract the best nEXP experts (with $\text{nEXP} \leq N_{\text{exp}}$) according to their strength in the linear fitting of $\hat{\mathbf{B}}_j$. In other words, we extract the indices $\{k_{j_1}, \dots, k_{j_{\text{nEXP}}}\}$ that maximize $\left\{ \sum_{l=1}^{\text{RED}} (X_{v_{j,l},j}^{(k)})^2, k = 1, \dots, N_{\text{exp}} \right\}$. The final prediction is given robustly as the median of the predictions of these experts $\hat{B}_{t_i,j} = \text{median}\{\hat{B}_{t_i,j}^{(k)}, k \in \{k_{j_1}, \dots, k_{j_{\text{nEXP}}}\}\}$. The Experts method here proposed is thus based on the set of parameters α, r (or, equivalently, *factor*), RED , nEXP , which must be tuned to balance the computational costs and the precision of the prediction.

4 Numerical results

The method has been applied to predict iteratively 12 time steps forward, for each input matrix B of the test set. The method was not much sensitive to α , thus we fixed $\alpha = 1$, and nEXP has been fixed to 11, since it was the maximum value which guaranteed us to remain in the maximum allowed execution time in the H2020 competition. We made the parameters *factor* and RED vary, and we computed the corresponding MSE, and the ratio $\text{MSE}(\text{expert method})/\text{MSE}(\text{constant prediction})$. The results are reported in the tables below.

	MSE						RATIO					
	RED	3	5	7	9	12	RED	3	5	7	9	12
factor							factor					
0.00001	935	900	875	876	944		0.00001	0.924827	0.890208	0.865480	0.866469	0.933729
0.00010	935	900	875	876	944		0.00010	0.924827	0.890208	0.865480	0.866469	0.933729
0.00100	933	901	874	875	944		0.00100	0.922849	0.891197	0.864491	0.865480	0.933729
0.01000	926	929	930	952	938		0.01000	0.915925	0.918892	0.919881	0.941642	0.927794
0.10000	1250	1250	1250	1250	1250		0.10000	1.236400	1.236400	1.236400	1.236400	1.236400

References

1. H2020 inducement prize: Big Data technologies 2017. <http://ec.europa.eu/research/horizonprize/index.cfm?prize=bigdata>
2. Sudipto, B., Anindya, R.: Linear Algebra and Matrix Analysis for Statistics, CRC Press, 2014.
3. Yanai, H., Takeuchi, K., Takane, Y.: Projection matrices, generalized inverse matrices, and singular value decomposition, Springer, 2011.

Regression with time-dependent PDE regularization for the analysis of spatio-temporal data

Regressione con regolarizzazione di PDE tempo dipendenti per modellizzare dati spazio-temporali

Eleonora Arnone, Laura Azzimonti, Fabio Nobile, Laura M. Sangalli

Abstract We propose a method for the analysis of space-time data when prior knowledge on the phenomenon under study is available. The model is based on regression with differential regularization. The differential regularization includes a time-dependent Partial Differential Equation (PDE) that formalizes problem-specific information on the phenomenon.

Abstract Proponiamo un metodo per l'analisi di dati spazio-temporali, quando è disponibile una conoscenza a priori sul fenomeno in esame. Il modello è basato su modelli di regressione con regolarizzazione differenziale. La regolarizzazione include un'equazione alle derivate parziale tempo dipendente che formalizza le informazioni sul fenomeno.

Key words: Finite Elements, Partial Differential Equation, Penalized regression, Smoothing

Eleonora Arnone

MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, e-mail: eleonora.arnone@polimi.it

Laura Azzimonti

IDSIA - Department of Innovative Technologies, Università della Svizzera Italiana Galleria 1, via Cantonale, Switzerland e-mail: laura.azzimonti@supsi.ch

Fabio Nobile

MATHICSE-CSQI, École Polytechnique Fédérale de Lausanne, Route Cantonale, 1015 Lausanne, Switzerland e-mail: fabio.nobile@epfl.ch

Laura M. Sangalli

MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy e-mail: laura.sangalli@polimi.it

1 Regression with time-dependent PDE Regularization

We propose a new method to model space-time data. The proposed method extends the models proposed in [8, 9, 4, 5] and is based on regression with differential regularization. The applications we are interested in are in particular the ones in which prior knowledge on the phenomenon under study is available. The prior knowledge is described in terms of a time-dependent PDE that jointly models the spatial and temporal variation of the phenomenon.

Methods based on differential regularizations has been recently developed in [3, 7, 6, 1]. The roughness penalty considered in these work are composed by two separated terms: one term accounts for the the regularity of the field in space, while the other term accounts for the regularity in time. Differently from the latter works, in [2] we propose a unique regularizing term that jointly model the space-time variation of the phenomenon, on the base of problem-specific prior information. Specifically, the regularization involves the misfit of a time-dependent parabolic PDE modeling the phenomenon behavior, $\dot{f} + Lf = u$, where \dot{f} is the time derivative of the spatio-temporal function f , and L is a differential operator in space of the form

$$Lf = -\text{div}(\mathbf{K}\nabla f) + \mathbf{b} \cdot \nabla f + cf$$

with $\mathbf{K} \in \mathbb{R}^{2 \times 2}$ a symmetric and positive definite diffusion tensor, $\mathbf{b} \in \mathbb{R}^2$ a transport vector and $c \geq 0$ a reaction term. We consider various samplings designs, including geo-statistical and areal data. We show that the corresponding estimation problems are well posed and can be discretized in space by means of the Finite Element method, similarly to [9] and [5], and in time by means of the Finite Difference method. The models can handle data distributed over domains having complex shapes and where the shape of the domain influences the phenomenon under study. Moreover, various types of boundary conditions can be considered, modelling the behavior of of the spatio-temporal field and the boundaries of the domain of interest.

2 Blood flow velocity field estimation in the carotid artery

This methodology was motivated by the study of the blood flow velocity field in the common carotid artery over a heart beat, starting from the Echo-Color Doppler (ECD) data analyzed for a single time instant in [5]. The echo doppler scan provides a time-dependent measure of the velocities of blood flow particles sampled within a bean in an artery section. Figure 2 shows the ECD signal registered in a centrally located bean at the cross-section of the common carotid artery located 2 cm before the artery bifurcation. The lower part of the ECD image displays the acquired velocity signal during the time lapse of about three heart beats. This signal represents the time-evolving histogram of the measured velocities in the beam: the gray-scaled intensity of pixels is proportional to the number of blood-cells in the beam moving at a certain velocity for any fixed time. The green line in the figure is the noisy

estimated mean velocity for each time instant. Starting from the ECD signals over multiple beams, we would like to reconstruct the time-varying mean velocity of the blood-flow.

The time-dependent PDE included in the regularization term for the estimation of the blood velocity is $\gamma \dot{f} + Lf = 0$ where the spatial operator L is the same derived in [5] for the estimation of the velocity at a fix time instant. The diffusion matrix \mathbf{K} is defined in

$$\mathbf{K}(x, y) = \begin{bmatrix} y^2 + \kappa_1 x^2 & (\kappa_1 - 1)xy \\ (\kappa_1 - 1)xy & x^2 + \kappa_1 y^2 \end{bmatrix} + \kappa_2 (R^2 - x^2 - y^2) \mathbf{I}_2,$$

with $R = 2.8$, $\kappa_1 = 0.1$, $\kappa_2 = 0.2$, $\mathbf{b}(x, y) = (\beta x, \beta y)'$ with $\beta = 0.5$. The relative strength between the space and time derivatives is controlled via the multiplying factor γ , which is set equal to 0.1. The estimated field is imposed to be zero on the boundary of the carotid artery, indeed a fundamental constraint in this application is given by the so-called no-slip boundary conditions. The physics of the problem implies that blood-flow velocity is zero at the arterial wall because of the friction between blood cells and arterial wall. The estimated dynamic surface is shown at fixed time instants in Figure 2. Including the prior knowledge about the blood fluid dynamics and appropriate boundaries conditions is in this context fundamental to achieve meaningful and physiological estimates.

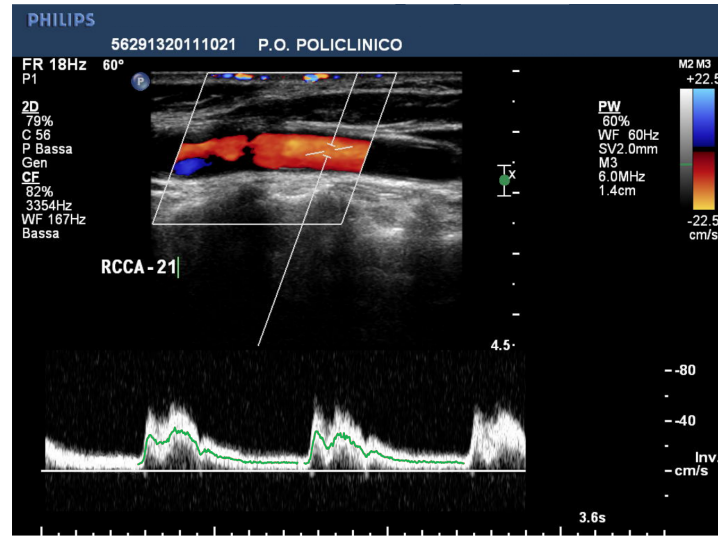


Fig. 1 ECD image corresponding to the central point of the carotid section located 2 cm before the carotid bifurcation, with the mean estimated velocity over two heart beats.

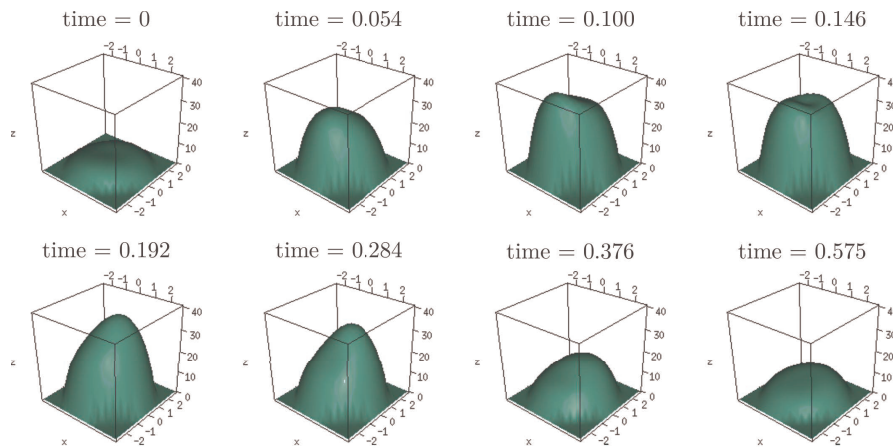


Fig. 2 Blood flow velocity field estimated by Space Time regression with time-dependent PDE regularization, at different time instants.

References

1. Aguilera-Morillo M. C., Durbán M., Aguilera A. M.: Prediction of functional data with spatial dependence: a penalized approach. *Stochastic Environ Res Risk Assess* **31**(1): 7–22 (2017)
2. Arnone E., Azzimonti L., Nobile F., Sangalli L. M.: Modeling spatially dependent functional data via regression with differential regularization. *Journal of Multivariate Analysis*, textbf170: 275–295 (2019)
3. Augustin N. H., Trenkel V. M., Wood S. N., Lorance P.: Space-time modelling of blue ling for fisheries stock management. *Environmetrics* **24**(2): 109–119 (2013)
4. Azzimonti L., Nobile F., Sangalli L. M., Secchi P.: Mixed Finite Elements for Spatial Regression with PDE Penalization. *SIAM/ASA Journal on Uncertainty Quantification* **2**(1): 305–335 (2014)
5. Azzimonti L., Sangalli L. M., Secchi P., Domanin M., Nobile F.: Blood flow velocity field estimation via spatial regression with PDE penalization. *Journal of the American Statistical Association* **110**(511): 1057–1071 (2015)
6. Bernardi M. S., Sangalli L. M., Mazza G., Ramsay J. O.: A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. *Stochastic Environmental Research and Risk Assessment*, **31**(1): 23–38 (2017)
7. Marra G., Miller D. L., Zanin L.: Modelling the spatiotemporal distribution of the incidence of resident foreign population. *Statistica Neerlandica* **66**(2): 133–160 (2012)
8. Ramsay, T.: Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **54**(2): 307–319 (2002)
9. Sangalli, L. M. and Ramsay, J. O. and Ramsay, T. O.: Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(4): 681–703 (2013)

A network analysis of museum preferences: the Firenzecard experience.

Un'analisi di rete delle preferenze museali: l'esperienza della Firenzecard

Silvia Bacci, Bruno Bertaccini, Roberto Dinelli, Antonio Giusti, and Alessandra Petrucci

Abstract Firenzecard is the official museum pass of the municipality of Florence that allows the visit of over eighty collections and exhibitions located in Florence and in the surrounding area. Each museum access is detected through apposite electronic devices; moreover, some characteristics of tourists (e.g., gender, age, citizenship) are recorded. Thus, the amount of data collected through Firenzecard is truly remarkable and represents an important source of information for improving services to tourists. In this contribution, we carry out a network analysis on the data collected in 2018 with the aim of exploring the sequences of museums visited and suggesting alternative visit opportunities.

Abstract *La Firenzecard rappresenta il pass ufficiale del Comune di Firenze per avere un accesso agevolato ad oltre ottanta musei (collezioni permanenti e mostre temporanee) di Firenze e dei comuni limitrofi. Ciascun accesso ai musei viene rilevato elettronicamente; in aggiunta, vengono registrate alcune caratteristiche dei turisti (ad es., sesso, età, cittadinanza). Tramite la Firenzecard viene raccolto un ingente ammontare di dati, che costituisce un'importante fonte di informazione per il miglioramento dei servizi turistici. In questo lavoro verrà svolta un'analisi di rete sui dati raccolti durante il 2018 al fine di esplorare le sequenze di musei visitati e suggerire percorsi di visita alternativi.*

¹ Silvia Bacci, Dipartimento di Statistica, Informatica, Applicazioni “G. Parenti”, Università di Firenze; email: silvia.bacci@unifi.it

Bruno Bertaccini, Dipartimento di Statistica, Informatica, Applicazioni “G. Parenti”, Università di Firenze; email: bruno.bertaccini@unifi.it

Roberto Dinelli, Linea Comune; email: r.dinelli@lineacomune.it

Antonio Giusti, Dipartimento di Statistica, Informatica, Applicazioni “G. Parenti”, Università di Firenze; email: ag@unifi.it

Alessandra Petrucci, Dipartimento di Statistica, Informatica, Applicazioni “G. Parenti”, Università di Firenze; email: alessandra.petrucci@unifi.it

Key words: Firenzecard, social network analysis, tourism

1 Introduction

Firenzecard is the official museum pass of the municipality of Florence. The card is valid for 72 hours starting from the first entrance and allows the tourist to visit all the affiliated museums in Florence and in the surrounding areas (about 80 permanent collections and temporary exhibitions). Each visiting experience, which cannot be repeated, does not require any reservation and allows the priority entry, with free admission for the children of the family. Furthermore, with a small surcharge, the card allows tourists to access the public transport system free of charge and the opportunity to take advantage of various offers from affiliated partners (mainly food and drinks or discounts for shopping).

Visits are automatically recorded at the entrance with optical reading devices. Therefore, each card provides lot of information such as number of museums visited, their sequence, and the time of entry. Furthermore, the cards sold online record information on gender, age, citizenship and family composition. Moreover, online cards are linked with a mobile app that administers a satisfaction questionnaire at the end of the visiting experience.

Since the start of the Firenzecard project in 2011, the total number of cards sold has reached approximately 720,000 units, of which 127,092 sold in the year 2018 (+4% compared to 2017). The average number of museums visited per card is 7.3.

The huge amount of data produced by the cards is a mine of potential information not yet fully explored. For example, data at issue provide useful information on the preferences of the tourist (in terms of sequences of museums visited), on the time spent in the individual visits, on the level of satisfaction associated with the experience, and on the overcrowding in real-time.

It is known that, in Florence, there are some very famous museums (e.g., Galleria degli Uffizi, Galleria dell'Accademia, Palazzo Pitti, Palazzo Vecchio or Palazzo della Signoria) which are almost always chosen by tourists, at the expense of those that are less known.

The purpose of this contribution is to analyse the sequences of the museums visited, to suggest alternative visiting opportunities. We intend to pursue this goal through the methods and techniques belonging to the Social Network Analysis (Kolaczyk, 2009; Luke, 2015); see also D'Agata et al. (2013) and Asero et al. (2016) for examples of network analysis applied to tourism data.

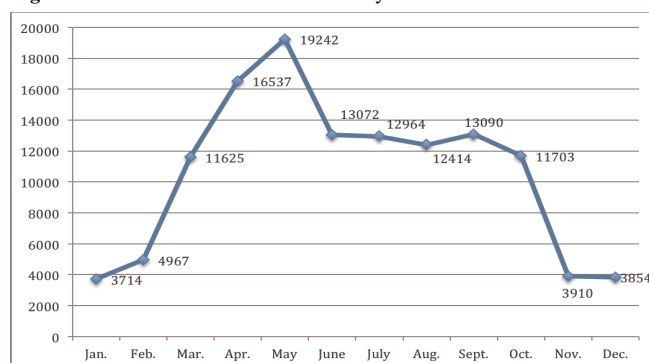
2 The dataset

In this contribution we focus on the data relating to the 127,092 cards sold in the year 2018 which correspond to a total of 884,389 visits to museums.

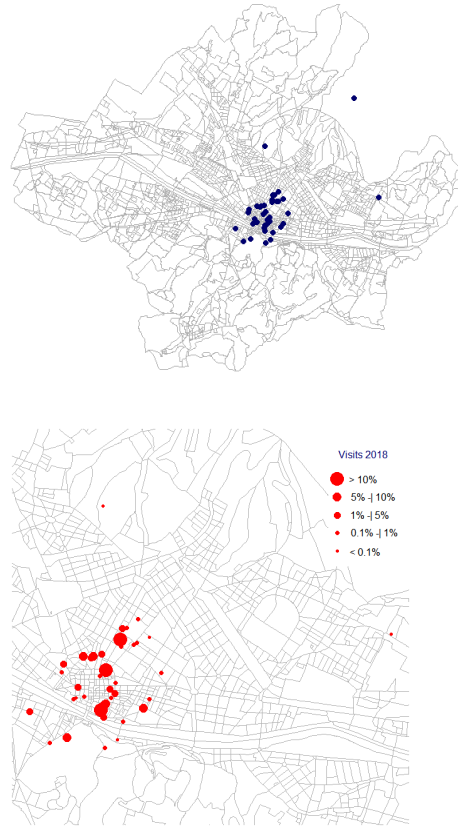
Firenzecard users are mainly aged between 25-44 years (60%) and 45-65 years (40%), with an average age of 44 years. Furthermore, 48.8% of users come from Europe (12% from France, followed by Italy with 11%), 37.5% from North America, and the 8.5% from Asia.

Figure 1 shows the trend of the cards sold monthly in 2018, and recalls the trend in the number of tourists in Florence during the year. We observe a substantial increase from January to May, followed by a sharp reduction (-32.1%) in June, due to an increase of the price of the cards. Then, card sales remain constant during the summer until October and are reduced again in November (-66.6% compared to the previous month) due to the reduced number of tourist presences in the last part of the year.

Figure 1: Number of Firenzecards sold in year 2018



As regards the typology of the museums visited, 45% of the visits are addressed to State museums, 13% to municipal museums, and the remaining visits interest other types of museums (e.g., private, university, churches and religious museums). As shown in Figure 2 (top panel), almost all the museums are in the city centre with differences in terms of visit rates (Figure 2 bottom panel). More in detail, the ranking of the top ten museums is shown in Table 1. Galleria degli Uffizi, Galleria dell'Accademia, and OPA (i.e., Brunelleschi's dome, Baptistery, Giotto's bell tower) collect over 10% museum visits; overall, the top ten museums collect 75% of all visits.

Figure 2: Map of museums (top panel) and focus on museums of the Municipality of Florence by rates of visit (bottom panel)**Table 1:** Visits per museum (number of visits and percentages)

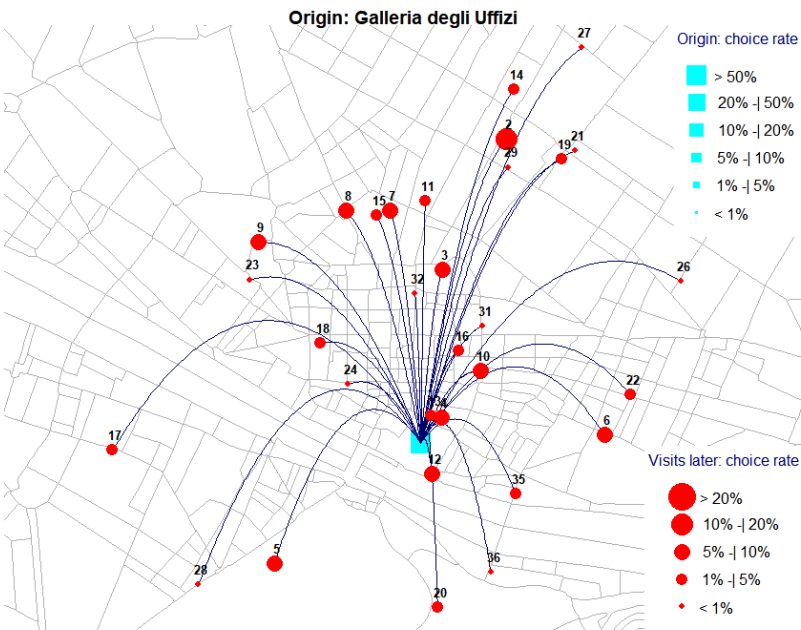
<i>ID</i>	<i>Museum</i>	<i>Visits</i>	<i>%</i>
1	Galleria degli Uffizi	104,864	11.9
2	Galleria dell'Accademia	100,691	11.4
3	OPA	94,744	10.7
4	Palazzo Vecchio	71,871	8.1
5	Palazzo Pitti	69,595	7.9
6	Santa Croce	54,722	6.2
7	San Lorenzo	47,980	5.4
8	Cappelle Medicee	44,714	5.1
9	Santa Maria Novella	44,456	5.0
10	Bargello	35,590	4.0
	Others	215,162	25.0

3 A preliminary network analysis

The sequences of the museums visited contribute to defining a network whose analysis is useful for identifying the most commonly chosen paths and, therefore, for providing tourists with suggestions on alternative models of visit.

Figure 3 shows the result of a preliminary network analysis on a random sample of 500 cards sold in 2018. The analysis focuses on the choice rates, conditionally on the previous visited museum. The square node indicates the origin of the network, that is, the previous visited museum; instead, the circular nodes denote the possible destinations of tourists. The size of the nodes is proportional to the choice rates. More in detail, the size of the square node shows the rate of Firenzecard users who visited the museum of origin during their stay in Florence. Instead, the size of the circular nodes is proportional to the percentage of tourists that visited a certain museum, immediately after the museum of origin.

Figure 3: Network from Galleria degli Uffizi (museum of origin) to other destinations



In detail, the example in Figure 3 shows a network originating from the Galleria degli Uffizi, visited by over 50% of Firenzecard users. After the Galleria degli Uffizi, the museum most often selected is the Galleria dell'Accademia (label 2 in the figure) that is chosen by 10-20% of tourists, followed, among others, by OPA (label 3), Palazzo Vecchio (label 4), Palazzo Pitti (label 5), and Santa Maria Novella (label 9), with a percentage of preferences between 5% and 10%. In contrast, Torre di Arnolfo di Cambio (label 13), which is very close to the Galleria degli Uffizi, is chosen by less than 5% of tourists.

For the future development of the work we intend to extend the analysis to all the sequences of visits (from the first to the last museum visited), including information on individual characteristics. Up to this point, we are developing probabilistic network models and related graphical tools, which adequately adapt the complexity of the network at issue.

References

1. Asero, V., Gozzo, S., Tomaselli, V.: Building tourism networks through tourist mobility. *J Travel Res* **55**, 751-763 (2016)
2. D'Agata, R., Gozzo, S., Tomaselli, V.: Network analysis approach to map tourism mobility. *Qual. Quant.* **47**, 3167-3184 (2013)
3. Kolaczyk, E. D.: *Statistical Analysis of Network Data. Methods and Models*. Springer, New York (2009)
4. Luke, D. A.: *A User's Guide to Network Analysis in R*. Springer, New York (2015)

A statistical learning approach to group response categories in questionnaires

Un approccio basato sull'apprendimento statistico per raggruppare le categorie di risposta nei questionari

Michela Battauz

Abstract Questionnaires still represent one of the main sources of data collection. The possibility of administering them online, eventually through mobile devices, has increased the amount of information collected. This requires the use of novel statistical tools, able to treat large amounts of data. In this paper we present a procedure to group the response categories of the items of a questionnaire, thus leading to a more parsimonious model. To this end, the nominal response model with a lasso-type penalty is employed. The proposal is illustrated through an application to data on the satisfaction with town services.

Abstract I questionari rappresentano ancora una delle principali fonti di raccolta di dati. La possibilità di somministrazione online, eventualmente attraverso dispositivi mobili, ha aumentato la quantità di informazione raccolta. Questo richiede l'uso di nuove tecniche statistiche, capaci di trattare grandi quantità di dati. In questo articolo presentiamo una procedura per raggruppare le categorie di risposta delle domande di un questionario, portando quindi a un modello più parsimonioso. A tal fine, si impiegherà il modello di risposta nominale con una penalizzazione di tipo lasso. La proposta è illustrata attraverso un'applicazione a dati sulla soddisfazione dei servizi offerti da una città.

Key words: fused lasso, item response theory, lasso, penalized likelihood, regularization.

1 Introduction

Item response theory (IRT) includes various statistical models for the analysis of the responses given to a test or questionnaire [1]. In these models, the probability of

Michela Battauz

Department of Economics and Statistics, University of Udine, via Tomadini 30/A Udine e-mail: michela.battauz@uniud.it

giving a certain response depends on a latent variable of interest and some parameters related to the items. Some IRT models are suited for the analysis of polytomous items. Among these, the graded response model [6] and the generalized partial credit model [4] assume that the response categories have a predetermined order. Instead, the nominal response model [2] does not require the ordering of the response options. However, it is not only used for modelling unordered responses, but it is also useful to check the expected order of the categories [7]. It is certainly the most flexible IRT model for polytomous items, but involves the estimation of many parameters that can be very unstable in small samples. In this paper, we propose the use of a lasso-type penalty [3, 9] to group the response categories and provide regularized estimates. The methodology will be presented in Section 2, and it will be illustrated through a real-data example in Section 3. Some concluding remarks will be given in Section 4.

2 Regularized estimation of the nominal response model

Suppose that item j has m_j possible responses, which are indicated with $k = 0, \dots, m_j - 1$. In the nominal response model, the probability of giving the response k to item j for subject i is given by:

$$P(Y_{ij} = k | \theta_i) = \frac{e^{\alpha_{jk}\theta_i + \beta_{jk}}}{\sum_{h=0}^{m_j-1} e^{\alpha_{jh}\theta_i + \beta_{jh}}}, \quad (1)$$

where θ_i represents a latent variable, while α_{jk} and β_{jk} are parameters. The slope parameters α_{jk} capture the relation between the latent variable and the probability of giving response k , and higher values of this parameter indicate that a certain response is more likely to be given by subjects with higher values of the latent variable. When the slope parameters of two response categories of the same item are equal, these can be collapsed [8]. Instead, the intercept parameters β_{jk} are related to the number of subject giving response k for item j . The parameters of the model are usually estimated by means of the marginal maximum likelihood method, maximizing the following log-likelihood function:

$$\ell(\alpha, \beta) = \sum_{i=1}^n \sum_{j=1}^J \log \int_{\mathbb{R}} \prod_{k=0}^{m_j-1} P(Y_{ij} = k | \theta_i)^{I(Y_{ij}=k)} \phi(\theta_i) d\theta_i, \quad (2)$$

where n is the number of subjects, J is the number of items, $I(\cdot)$ is an indicator function, and $\phi(\cdot)$ is the density of a standard normal variable. In order to achieve a more parsimonious model and to limit the variability of the parameter estimates, the proposal is to include a penalty term in the log-likelihood function that forces the slope parameters of the same item to assume the same value. The penalized log-likelihood function is as follows:

$$\ell_p(\alpha, \beta) = \ell(\alpha, \beta) - \lambda \sum_{j=1}^J \sum_{k=0}^{m_j-2} \sum_{h=k+1}^{m_j-1} |\alpha_{jk} - \alpha_{jh}|. \quad (3)$$

The penalty in (3) is similar to a fused-lasso penalty [10]. However, in this case, there is not a natural order of the parameters, and hence all the pairs of slope parameters of the same item should be considered. Similarly to [11], we explored also a adaptive version of the penalty that includes weights that depend on the data:

$$\ell_p(\alpha, \beta) = \ell(\alpha, \beta) - \lambda \sum_{j=1}^J \sum_{k=0}^{m_j-2} \sum_{h=k+1}^{m_j-1} |\alpha_{jk} - \alpha_{jh}| w_{jkh}, \quad (4)$$

with

$$w_{jkh} = |\hat{\alpha}_{jk}^{MLE} - \hat{\alpha}_{jh}^{MLE}|^{-1}, \quad (5)$$

where $\hat{\alpha}_{jk}^{MLE}$ denotes the maximum likelihood estimate. When the maximum likelihood estimates are very close to each other, these weights become very high thus encouraging a fusion between the parameters.

3 A real-data example

The proposed methodology was applied to the 2015 Chapel Hill Community Survey¹, designed to investigate the satisfaction of the residents with the town services. The questionnaire is composed of 18 items with 5 possible response options, which are: *very dissatisfied*, *dissatisfied*, *neutral*, *satisfied*, *very satisfied*. The sample size is equal to 407. The survey explores the satisfaction of the residents with respect to a variety of aspects of the public services including, for example, safety, parks and recreation programs, library services, maintenance of streets, maintenance of buildings, or flow of traffic. Overall, the satisfaction with the public services is quite good: the median is equal to *satisfied* for most of the items; it is equal to *neutral* for only 4 items and it is equal to *very satisfied* for 1 item.

All analyses were performed with the R software [5], employing the `mirt` package for the estimation of the nominal model with the maximum likelihood method, and functions written by the authors to implement the maximum penalized likelihood estimation proposed in this paper. The value of the tuning parameter λ was selected through 5-fold cross-validation, and it is represented by the vertical dotted line in Figure 1. The Figure shows the regularization path of two items taken as example, where the parameter estimates are plotted against increasing values of λ . Category *very dissatisfied* is not represented in the graph because it was the reference category, with coefficients set to zero to assure the identification of the parameters. Using the non-adaptive penalization, only the slope parameter of category *dissatisfied* of the item on the left panel is fused with the slope parameter of category

¹ <https://catalog.data.gov/dataset/community-survey-q1-overall-satisfaction-with-town-services>

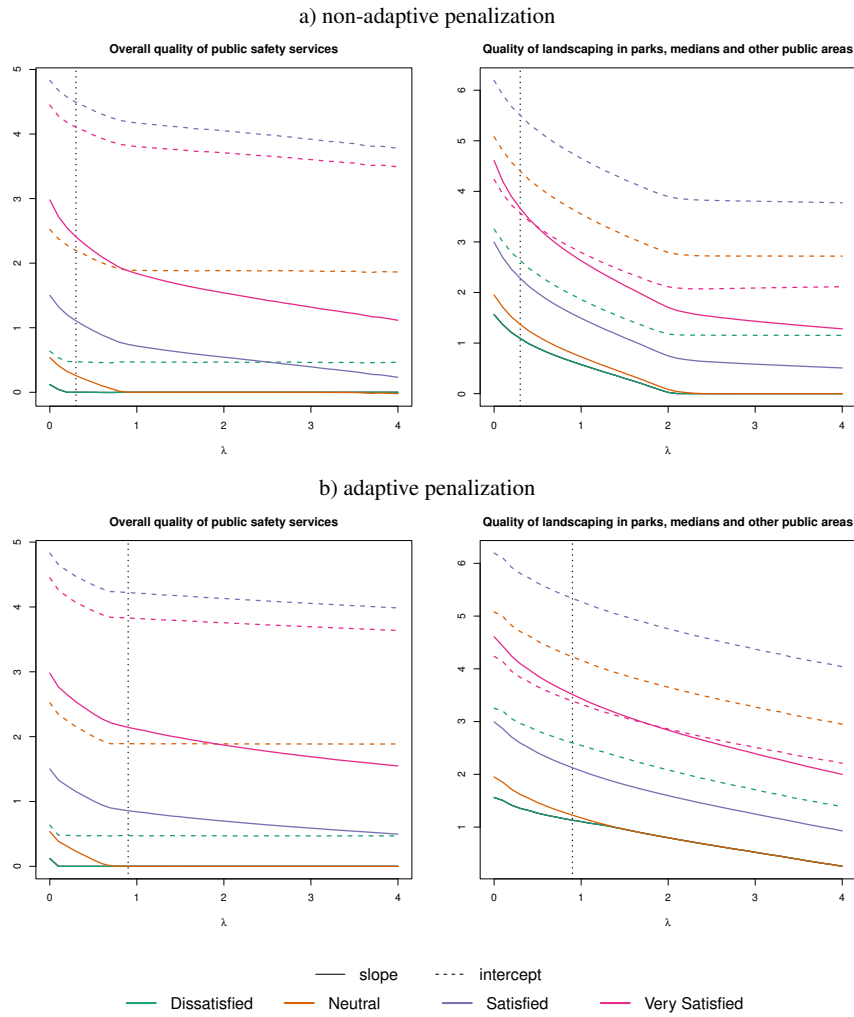


Fig. 1 Regularization path of two items.

very dissatisfied at the selected value of λ . Applying the adaptive penalization, the slope parameters of both categories *dissatisfied* and *neutral* are fused with the slope parameter of category *very dissatisfied* at the selected value of λ for the item on the left panel. Hence, these three categories of this item can be collapsed. The item on the right panel does not present slope parameters fused at the selected value of λ . However, all the parameters present shrunk values.

Figures 2 and 3 show the probability curves for the same two items when $\lambda = 0$, i.e. using the maximum likelihood estimates, and when $\lambda = 0.9$, which is the value selected by cross-validation using the adaptive penalization.

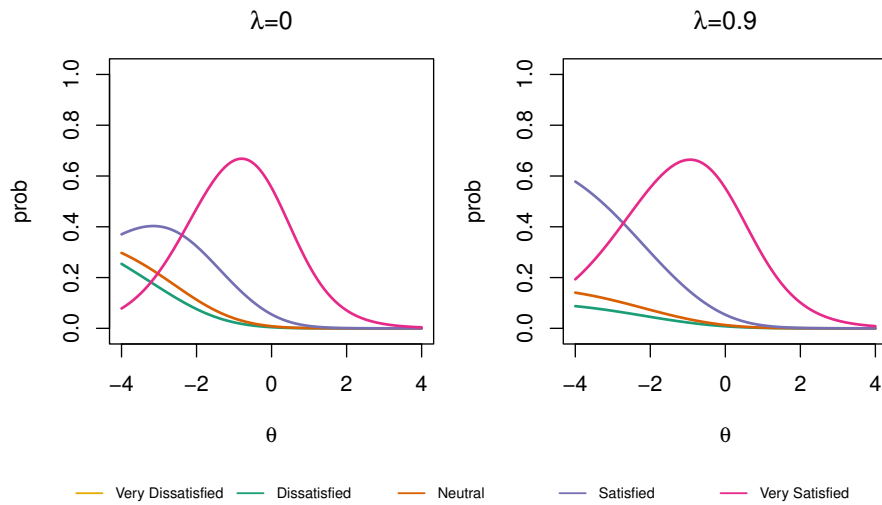


Fig. 2 Probability curves of item *Overall quality of public safety services*.

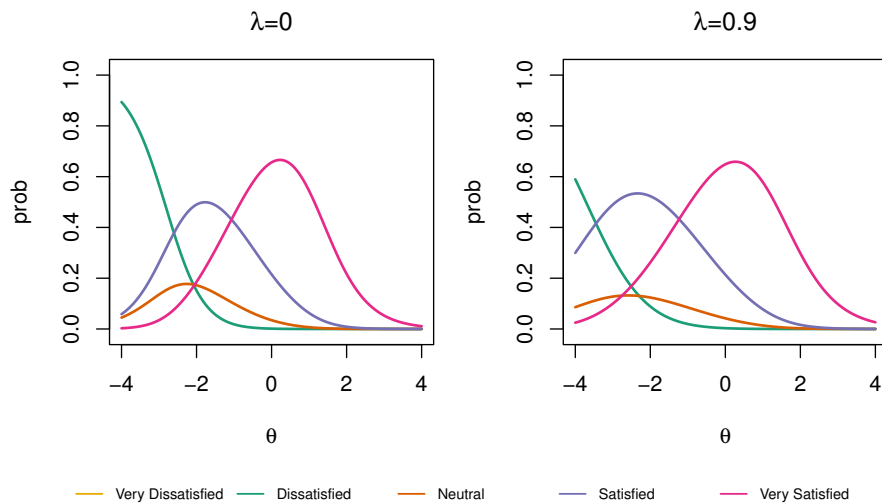


Fig. 3 Probability curves of item *Quality of landscaping in parks, medians and other public areas*.

4 Conclusions

The procedure proposed is quite effective in grouping the response categories that present a similar relation with the latent variable, especially using the adaptive version of the penalty. These categories can be collapsed, thus leading to a more parsimonious model. Since this procedure is meant to remove the noise eventually

present in the data, it does not lead to a loss of information when the response categories are grouped. In this respect, it is very important the selection of the tuning parameter λ , and cross-validation is a method suited to select a model that describes the data generating process. Besides of indicating the categories that can be collapsed, the method provides also regularized estimates and hence it represents a promising method for improving the efficiency of the estimators. A simulation study to better understand the performance of the method is under study.

Acknowledgements This work was partially supported by PRIN 2015 prot. 2015EASZFS.003 and partially by PRID 2017, University of Udine.

References

1. Bartolucci, F., Bacci, S., Gnaldi, M.: Statistical analysis of questionnaires: A unified approach based on R and Stata. Chapman and Hall/CRC, Boca Raton (2015)
2. Bock, D. R.: Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29–51 (1972)
3. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC, New York (2015)
4. Muraki, E.: A generalized partial credit model: Application of an EM algorithm. *Appl. Psychol. Meas.*, **16**, 159–176 (1992)
5. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018)
6. Samejima, F.: Estimation of ability using a response pattern of graded scores. *Psychometrika*, **34**, 1–97 (1969)
7. Thissen, D., Cai, L., & Bock, R. D.: The nominal categories item response model. In: Nering, M. L., Ostini, R. (eds.) *Handbook of Polytomous Item Response Theory Models*, pp. 43–75. Routledge (2010).
8. Thissen, D. & Cai, L.: Nominal categories models. In: van der Linden, W. J. (ed.) *Handbook of Item Response Theory, Volume One: Models*, pp. 51–73. Chapman and Hall/CRC (2016).
9. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., B*, **58**, 267–288 (1996)
10. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K.: Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc., B*, **67**, 91–108 (2005)
11. Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.*, **101**, 1418–1429 (2006)

Tree-based Functional Data Analysis for Classification and Regression

Alberi di Classificazione e Regressione per dati Funzionali

Edoardo Belli, Enrico Ragaini, Simone Vantini

Abstract In this talk we will introduce a general and interpretable tree-based framework for classification and regression problems with inputs and/or outputs that belong to a separable functional Hilbert space. This approach is based on the use of a node-wise weight function which can be regularized to affect the decisions. We will show how the different embedding spaces of the functions have an impact both on interpretability and predictive power, with the possibility of using a non parametric approach for choosing the embedding space. This framework will also be extended to a Random Forest setting for additional prediction capabilities. Finally, the methods will be evaluated on a case study that belongs to the electrical engineering sector.

Abstract *Nella relazione si introdurrà un framework generale ed interpretabile, basato su alberi, per problemi di classificazione e regressione dove sia gli input che gli output possono essere funzioni che appartengono ad uno spazio funzionale di Hilbert separabile. Lo spazio in cui si immergono le funzioni ha un ruolo fondamentale sia per l'interpretabilità che per il potere predittivo. In particolare, la scelta dello spazio verrà fatta con un approccio non parametrico. Inoltre si estenderà questo framework ad un contesto di tipo Random Forest, con l'obiettivo di aumentare le performance predittive. Infine, verrà presentato un caso di studio che appartiene al settore dell'ingegneria elettrica.*

Edoardo Belli
Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy.
e-mail: edoardo.belli@polimi.it

Enrico Ragaini
ABB - Electrification Products, Via Pescaria 5, 24123, Bergamo, Italy.
e-mail: enrico.ragaini@it.abb.com

Simone Vantini
Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy.
e-mail: simone.vantini@polimi.it

Key words: Functional Data Analysis, Object Oriented Data Analysis, Classification, Regression, CART, Random Forest

1 Motivation

In recent years, Object Oriented Data Analysis (OODA) [10] has become a lively research topic, as data becomes more complex and new tools are needed to extract meaningful and reliable information from it. Functional Data Analysis (FDA) [15] can be seen as a data modeling framework for smooth and high dimensional data, which often arises in applications where high-frequency sampling devices are used [19] [18]. The FDA approach has been studied from the theoretical point of view [8] [11] and has been proposed for solving problems that deal with inference [14] [22] [7], clustering [20] [5], classification [9] [6] [1] [17] [2], regression [13] and time series analysis [3], both in the parametric and non parametric settings [4] [16]. While much attention has been given to parametric models like the functional linear model, functional tree-based methods [12] [21] are still an open topic that needs further investigation, as there are many relevant nonlinear problems which have proven to be complex from both the numerical and computational point of view.

2 Outline

In this talk we will present the problem of non parametric classification and regression with inputs and/or outputs that belong to a separable functional Hilbert space, with a specific focus on tree-based methods. After briefly recalling the multivariate Classification and Regression Trees (CART) and Random Forest algorithms, we will describe the current state of the art regarding Functional Data analysis in the context of tree-based methods. In fact, we will show that in the functional setting, all the tree-based methods that are found in the literature either deal with problems of classification or regression, lacking a general formulation. Moreover, these methods are restricted to functional inputs or outputs only. Therefore, we propose a general and interpretable tree-based framework for function (or scalar) on function regression and classification with functional predictors, which uses a weight function in each node of the tree. For each node, we solve the following optimization problem:

$$\begin{aligned} \min_{w,s} [\mathcal{L}(\mathcal{R}_1(w,s)) + \mathcal{L}(\mathcal{R}_2(w,s)) + \lambda \text{Pen}(w)] \quad (1) \\ \text{s.t. } \int w(t)dt = 1 \\ w(t) \geq 0 \end{aligned}$$

where \mathcal{L} is the loss function of the node which depends on the problem and $\text{Pen}(w)$ is a regularization term for the weight function. This formulation is independent from the definition of the splitting rule, which determines whether the input

function $x(t)$ will belong to \mathcal{R}_1 or \mathcal{R}_2 , the two regions that determine the splitting of the input space at the current node. We propose as a functional splitting criteria the weighted integral mean of the function:

$$\begin{aligned}\mathcal{R}_1(w, s) &= \{x(t) \mid \frac{1}{T} \int_I w(t)x(t)dt \leq s\} \\ \mathcal{R}_2(w, s) &= \{x(t) \mid \frac{1}{T} \int_I w(t)x(t)dt > s\}\end{aligned}$$

In particular, this weight function determines the embedding space of the input functions and can be used for regularization, imposing sparsity and to aid interpretability. The weight function plays a crucial role in deciding the splitting feature and threshold of each node, which is a non trivial optimization problem that does not have a closed form solution. Therefore we will discuss possible heuristics based on local function approximations and other global optimization algorithms. This new type of tree will be then extended to a Random Forest context, where the weighting function takes a crucial role in the randomization of the decisions. The methods will be evaluated with simulated data, taking the functional linear model as a baseline reference. Finally, we will present a case study that belongs to the electrical engineering sector and in particular is related to quality control for circuit breakers.

Acknowledgements The case study with the corresponding data is provided by ABB through the joint research center with Politecnico di Milano and the research funded by the PhD scholarship "Development and prototyping of distributed control systems for electric networks based on advanced statistical models for the analysis of complex data".

References

1. Alonso A.M., Casado D., Romo J.: Supervised classification for functional data: A weighted distance approach. *Computational Statistics and Data Analysis*. 56 2334-2346 (2012)
2. Bin L., Yu Q.: Classification of functional data: A segmentation approach. *Computational Statistics and Data Analysis*. 52 4790-4800 (2008)
3. Canale A., Vantini S.: Constrained functional time series: Applications to the Italian gas market. *International Journal of Forecasting*. 32 1340-1351. (2016)
4. Ferraty F., Vieu P.: *Nonparametric functional data analysis: theory and practice*. Springer, New York. (2006)
5. Floriello D., Vitelli V.: Sparse clustering of functional data. *Journal of Multivariate Analysis*. 154 1-17 (2018).
6. Galeano P., Joseph E., Lillo R.: The Mahalanobis Distance for Functional Data with Applications to Classification. *Technometrics*. 57:2 281-291 (2015)
7. Horváth L., Kokoszka P.: *Inference for functional data with applications*. Springer, New York. (2012)
8. Hsing T., Eubank R.: *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons. (2015)
9. James G. M., Hastie T. J.: Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*. 63 533-550 (2001)
10. Marron J. S., Alonso A. M.: Overview of object oriented data analysis. *Biometrical Journal*. (2014)

11. Menafoglio A., Secchi P.: Statistical analysis of complex and spatially dependent data: A review of object oriented spatial statistics. *European Journal of Operational Research*. (2016)
12. Moller A., Tutz G., Gertheiss J.: Random Forests for functional covariates. *Journal of Chemometrics*. 30:715-725. (2016)
13. Muller H.G., Stadtmuller U.: Generalized functional linear models. *Annals of Statistics*. 774-805. (2005)
14. Pini A., Stamm A., Vantini S.: Hotelling's T2 in separable Hilbert spaces. *Journal of Multivariate Analysis*. 167 284-305. (2018)
15. Ramsay J.O., Silverman B.W.: *Functional Data Analysis*. Springer. (2005)
16. Rossi F., Conan-Guez B.: Functional multi-layer perceptron: a non-linear tool for functional data analysis. *Neural Networks*. 18 45-60. (2005)
17. Rossi F., Villa N.: Support vector machine for functional data classification. *Neurocomputing*. 69 730-742. (2006)
18. Sangalli L.M., Secchi P., Vantini S., Veneziani A.: Efficient estimation of three-dimensional curves and their derivatives by free-knot regression splines, applied to the analysis of inner carotid artery centrelines. *Journal of the Royal Statistical Society, Series C: Applied Statistics*. 58 285-306. (2009)
19. Sangalli L.M., Secchi P., Vantini S., Veneziani A.: A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery. *Journal of the American Statistical Association*. 104:485, 37-48. (2009)
20. Sangalli L.M., Secchi P., Vantini S., Vitelli V.: k-mean alignment for curve clustering. *Computational Statistics and Data Analysis*. 54 1219-1233. (2010)
21. Yu Y., Lambert D.: Fitting Trees to Functional Data, with an Application to Time-of-Day Patterns. *Journal of Computational and Graphical Statistics*. 8:4 749-762. (1999)
22. Secchi P., Stamm A., Vantini S.: Inference for the mean of large p small n data: a finite-sample high-dimensional generalization of Hotellings theorem. *Electronic Journal of Statistics*. 7 20052031 (2013)

PDE-regularized regression for anisotropic spatial fields

Regressione con regolarizzazione differenziale per campi spaziali anisotropi

Mara S. Bernardi, Michelle Carey, James O. Ramsay and Laura M. Sangalli

Abstract We present a class of models for the analysis of spatially distributed data with complex dependencies. In particular, we consider data characterized by stationary spatial anisotropy. The penalized spatial regression method we propose makes use of an anisotropic diffusion operator in the regularization in order to properly take into account the spatial anisotropy present in the data. We consider the diffusion operator as unknown and we estimate it from the data. To this end, we adopt a profiling estimation approach, adapting to our setting the parameter cascading technique. The method can be extended to the case of non-stationary anisotropy.

Abstract *Consideriamo una classe di modelli per l'analisi di dati spazialmente distribuiti con dipendenze complesse. In particolare, consideriamo dati caratterizzati da anisotropia spaziale stazionaria. Il metodo di regressione spaziale penalizzata che proponiamo sfrutta un operatore di diffusione anisotropo nella regolarizzazione in modo da tenere in considerazione l'anisotropia spaziale presente nei dati. Consideriamo l'operatore di diffusione come incognito e lo stimiamo a partire dai dati. A tal fine, adattiamo al nostro contesto la tecnica denominata "parameter cascading". Il metodo può essere esteso al caso di anisotropia non stazionaria.*

Key words: Functional data analysis, Penalized regression, Anisotropy

Mara S. Bernardi

MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy

e-mail: marasabina.bernardi@polimi.it

Michelle Carey

School of Mathematics and Statistics, University College Dublin, Ireland

James O. Ramsay

Department of Psychology, McGill University, Montréal (Québec) Canada

Laura M. Sangalli

MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy

1 Introduction

Data arising from spatially distributed phenomena are often characterized by complex dependencies. In particular, we here consider the case of data showing spatial anisotropy. In biology, anisotropy is naturally induced by the arrangement and orientation of fibers and cells in a tissue, or by the morphology of the organs; in meteorology, it may be caused by the presence of winds and sea streams, or by the orography of the region under study; in geology, by the process of sedimentation.

Here we adopt a functional data analysis approach and propose to model spatial anisotropy via regression with partial differential regularization. [Ramsay \(2002\)](#) and [Sangalli et al. \(2013\)](#) consider spatial regression with a roughness penalty that involves the Laplacian of the spatial field: this partial differential operator provides a simple and isotropic measure of the curvature of the spatial field, and its use in the regularizing term induces an isotropic smoothing effect. We here extend the method to the case where the regularizing term involves a more general partial differential equation (PDE) that induces an anisotropic smoothing.

We use the PDE in the regularizing term to model the spatial variation of the phenomenon, learning the anisotropy from the data. Specifically, the PDE in the regularizing term induces a stationary anisotropic smoothing effect; the parameters in the PDE determine the direction and the intensity of the anisotropy; these parameters are here considered unknown and are estimated from data.

As in [Ramsay \(2002\)](#), [Sangalli et al. \(2013\)](#) and [Azzimonti et al. \(2015\)](#), we here represent the spatial field via finite elements, which provide a basis for piecewise polynomial surfaces.

Moreover, we estimate the parameters of the PDE in the regularizing term by parameter cascading. This generalized profiling estimation procedure was originally introduced by ([Ramsay et al., 2007](#)) to retrieve the parameters of an ordinary differential equation (ODE), starting from noisy measurements of the ODE solution. This technique has been successfully applied in other contexts. In our case, we use parameter cascading to obtain the parameters of a PDE in the regularizing term to characterize the spatial distribution of the data.

2 Model

Let $\{p_1 = (x_1, y_1), \dots, p_n = (x_n, y_n)\}$ be a set of n points on a bounded domain $\Omega \in \mathbb{R}^2$, whose boundary $\partial\Omega$ is a curve of class \mathcal{C}^2 . We denote by $|\Omega|$ the area of the domain Ω . Let $z_i \in \mathbb{R}$ be the value of a variable of interest observed at point p_i . We assume that z_1, \dots, z_n are noisy observations of an underlying smooth function $f : \Omega \rightarrow \mathbb{R}$. That is, for all $i \in \{1, \dots, n\}$,

$$z_i = f(p_i) + \varepsilon_i \quad (1)$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independently distributed residuals, with mean zero and constant variance σ^2 .

We want to estimate the spatial field f by minimizing the penalized sum-of-square-error functional

$$J_\rho(f, K) = (1 - \rho) \frac{1}{n} \sum_{i=1}^n \{z_i - f(p_i)\}^2 + \rho \frac{1}{|\Omega|} \int_{\Omega} \{\nabla \cdot (K \nabla f)\}^2, \quad (2)$$

where the operator ∇ is defined as $\nabla = (\partial/\partial x, \partial/\partial y)^\top$, K is a symmetric and positive definite matrix and $\rho \in (0, 1)$ is the smoothing parameter, which weighs the contribution of the data fitting term against the regularization term. The inclusion in the regularizing term of the anisotropic diffusion operator provides an anisotropic smoothing effect, where the direction and intensity of the anisotropy is determined by the matrix K .

We solve the non-convex optimization problem of minimizing the functional $J_\rho(f, K)$ with respect to f and K , with a value of ρ chosen to properly weigh the effect of the regularization and to accurately identify the optimal anisotropy matrix K .

We propose a two-step algorithm which alternates between the optimal selection of the anisotropy matrix K and the estimation of spatial field f . In the first step of the algorithm, the optimal anisotropy matrix K is selected using a parameter cascading approach. In the second step of the algorithm, the spatial field f is estimated using the anisotropy matrix K selected in the previous step of the algorithm. Once the penalization term is fixed, the estimator of the spatial field f exists and is unique under suitable assumptions.

For details on the implementation, see [Bernardi et al. \(2018\)](#).

3 Extension to non-stationary anisotropy

The proposed technique can be extended to the case where the anisotropy matrix K is spatially varying. This extension allows to analyze data characterized by non-stationary anisotropy.

In practice, we want to estimate the spatial field f by minimizing the following penalized sum-of-square-error functional:

$$J_\rho(f, K) = (1 - \rho) \frac{1}{n} \sum_{i=1}^n \{z_i - f(p_i)\}^2 + \rho \frac{1}{|\Omega|} \int_{\Omega} \{\nabla \cdot (K(\mathbf{p}) \nabla f)\}^2, \quad (3)$$

which is analogous to the functional (??), but where the matrix K is replaced by $K(\mathbf{p})$, a function defined on Ω taking values in the space of symmetric and positive definite 2×2 -matrices. $K(\mathbf{p})$ defines the non-stationary spatial anisotropy characterizing the differential operator $\nabla \cdot K(\mathbf{p}) \nabla f$.

This approach is similar to [Azzimonti et al. \(2015\)](#), where the method in [Ramsay \(2002\)](#) and [Sangalli et al. \(2013\)](#) is extended to the case where the regularizing term involves a more general linear second-order PDE with space-varying terms, which induces an anisotropic and non-stationary smoothing. In particular, [Azzimonti et al. \(2015\)](#) consider the case where prior knowledge about the physical phenomenon generating the data is available and can be formalized in terms of a PDE governing the phenomenon under study. Moreover, [Azzimonti et al. \(2015\)](#) assume that the problem-specific knowledge fully specifies the PDE in the regularizing term.

Like [Azzimonti et al. \(2015\)](#), we consider the case where the non-stationary spatial anisotropy is induced by the physics of the phenomenon under study. Differently from [Azzimonti et al. \(2015\)](#), we here instead consider the more general case where the prior knowledge is sufficient to specify the anisotropic term $K(\mathbf{p}, \theta)$ in (3) only up to some unknown parameters θ . We thus propose to estimate θ from the data, following the approach described in Section 2.

For example, the prior knowledge on the phenomenon may give information about the direction of the anisotropy, but its intensity should be learnt from the data.

References

- Azzimonti, L., Sangalli, L. M., Secchi, P., Domanin, M., and Nobile, F. (2015). Blood flow velocity field estimation via spatial regression with PDE penalization. *Journal of the American Statistical Association*, 110(511):1057–1071.
- Bernardi, M. S., Carey, M., Ramsay, J. O., and Sangalli, L. M. (2018). Modeling spatial anisotropy via regression with partial differential regularization. *Journal of Multivariate Analysis*, 167:15–30.
- Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796.
- Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):307–319.
- Sangalli, L. M., Ramsay, J. O., and Ramsay, T. O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):681–703.

A Bayesian model for network flow data: an application to BikeMi trips

Giulia Bissoli, Celeste Principi, Gian Matteo Rinaldi, Mario Beraha and
Alessandra Guglielmi

Abstract We propose a Bayesian model for the analysis of flow counts on a network for an application to a bike sharing platform (BikeMi) in Milano. Incorporating edge-specific covariates, we assume a zero-inflated Poisson mixture regression model, which easily accommodates for the sparse nature of the network under investigation.

Abstract *In questo lavoro proponiamo un modello bayesiano per l'analisi dei conteggi dei flussi su una rete, per un'applicazione relativa ad una piattaforma di bike sharing di Milano (BikeMi). Incorporando covariate specifiche ad ogni arco, assumiamo un modello mistura di Poisson zero-inflated, che permette di catturare facilmente la natura sparsa delle rete presa in considerazione.*

1 Introduction

Self service bike sharing systems have grown in popularity all over the world in recent decades, with dozens of new players entering the market each year. The exploitation of this novel transport system has generated an enormous quantity of data that could lead to a better understanding of city mobility. In this work, we analyze data from BikeMi, Milan oldest bike sharing system. BikeMi has been introduced in 2008, and nowadays encompasses more that 4000 bicycles with more than 250 stations.

Giulia Bissoli, Celeste Principi, Gian Matteo Rinaldi
Dipartimento di Matematica, Politecnico di Milano,
e-mail: {giulia.bissoli, celeste.principi, gianmatteo.rinaldi}@mail.polimi.it

Mario Beraha[†], Alessandra Guglielmi
Dipartimento di Matematica, Politecnico di Milano
e-mail: {mario.beraha, alessandra.guglielmi}@polimi.it

[†] Also affiliated with Università degli Studi di Bologna

We focus on the analysis of the flow of bikes from one station to another. By looking at the bike sharing system as a complex network [6], we assume a Bayesian regression model for the flow counts on the edges of the network. In particular, we assume a mixture model [5] for the flow counts, which yields a cluster estimate of the flows that we are able to interpret.

The structure of the paper is as follows: after having introduced the model in Section 2 and presented the dataset in Section 3, we report some posterior inference in Section 4, showing the goodness of fit of our model and the interesting insights from the estimated clusters. Finally in Section 5 we compare our model to competitors.

2 Zero-inflated Poisson mixture regression models

Let $\mathcal{G} = (V, E)$ be a directed graph with vertices V and edges $E = \{(i, j)\}$. In our application, we consider the problem of modelling the data flow on the edges E . For each edge (i, j) we define the variable Y_{ij} as the amount of bike trips from i to j taken on a fixed period of time. We also assume the availability of a set of covariates \mathbf{x}_{ij} for each arc, which, a priori, might influence the flow on that particular arc. The Y_{ij} s are thus counts, so that, a standard choice to model these variables would be then to use the Poisson distribution with suitable parameters. We also aim at capturing the topology of the network, i.e. assigning zero flow to the edges that should not be in the network; in addition, our model should be flexible enough to represent a wide class of scenarios, from routes travelled very often to the ones seldom used.

Combining these insights together led us to consider the following zero-inflated Poisson mixture regression model

$$Y_{ij} | \theta, \boldsymbol{\mu}_{ij}, \boldsymbol{\lambda} \stackrel{\text{ind}}{\sim} \begin{cases} \theta + (1 - \theta) \text{PM}(0 | \boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) & \text{if } Y_{ij} = 0 \\ (1 - \theta) \text{PM}(Y_{ij} | \boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) & \text{if } Y_{ij} > 0 \end{cases} \quad (1)$$

$$\log \mu_{ijk} = \boldsymbol{\beta}_k \mathbf{x}_{ij}, \quad (2)$$

for any couple of nodes (i, j) . We assume that, conditionally to all parameters, the Y_{ij} s are independent. Here PM stands for Poisson Mixture, i.e. $\text{PM}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_{i=1}^K \lambda_i \text{Poi}(\mu_i)$ with $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$, $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$, and K is a fixed positive integer. Observe that $\boldsymbol{\mu}_{ij} = (\mu_{ij1}, \dots, \mu_{ijK})$ is linked to the p -dimensional vector \mathbf{x}_{ij} via the canonical link function (2) through a p -dimensional regression parameter $\boldsymbol{\beta}_k$.

The prior specification assumes parameters $(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)$, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ independent with marginal distributions as follows:

$$\boldsymbol{\beta}_k \stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{m}, \Sigma) \quad k = 1, \dots, K \quad (3)$$

$$\boldsymbol{\lambda} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K) \quad (4)$$

$$\theta \sim \mathcal{U}(0, 1) \quad (5)$$

Assuming zero-inflated likelihood thus accounts for the presence of edges in the graph with zero flow, while the Poisson mixture let us easily model a various range

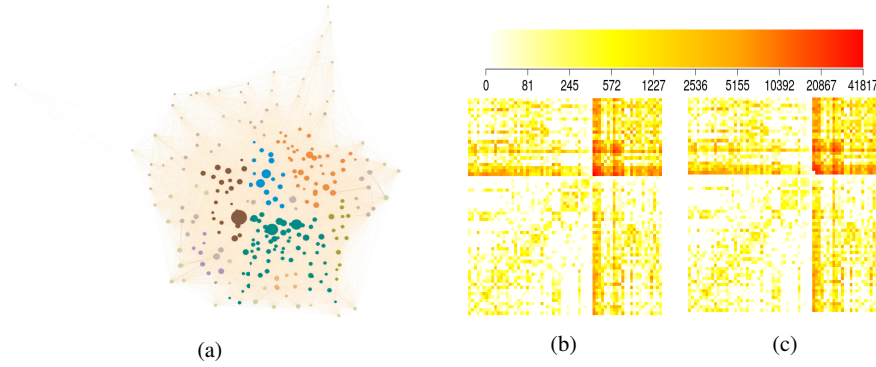


Fig. 1: (a): The nodes of the clustered network: the size of each dot is proportional to the number of stations it represents. (b): observed flow. (c): predicted flow. The entry of the matrix in row i , column j represents the flow on edge (i, j) .

of behaviours in the flows. Moreover, we take into account covariates as regressors in the generalized linear model through the locations μ_{ijk} .

3 BikeMi Dataset

We consider data arising from the most popular and oldest bike sharing system in Milan. The system is composed of 263 stations where users go and pick up and then drop off the bikes. The dataset consists of the records of 350,093 trips between pairs of stations between January 25th and March 6th, 2016. Using the stations as nodes in the graph would, in principle, lead to a graph with $263^2 \approx 69k$ edges, which makes the computational burden unfeasible.

As many stations lie within less than 100m from each others, we perform a clustering on the geo-locations of the stations, using the popular-density based clustering algorithm DBSCAN [4], and have consider the barycenter of each cluster as a node in the graph. In the end, we were left with 67 nodes. We report the clusters obtained using DBSCAN in Figure 1(a), which defines the nodes of the graph we analyze. We have then redefined y_{ij} , the observed flow counts between nodes i and j as the total flow from all the stations which were collapsed through DBSCAN procedure into node i and into node j .

Of course, other strategies are possible to reduce the dimensionality of the problem. For example one could consider Milan's neighborhoods (NILs) as nodes in the graph, as done in [7], and do not rely on the spatial clustering. This approach, however, would imply that trips from one node to itself (within the same NIL) might be longer than trips from one node to another, as two bike stations might be at the opposite sides of a NIL (former case) or just across the border of the two neighboring NILs (latter case). On the contrary DBSCAN clusters together points based on their local density, so that the estimated clusters can genuinely be well represented by their barycenters.

Similarly as in [3], we assume that bike flow from i to j might depend on: (i) the geographical distance between the stations (d_{ij}) (ii) the outer strength of the pick up node (S_i) and (iii) the inner strength of the drop off node (T_j). The outer (inner) strength of a node is defined as the total amount of trips departing from (arriving at) it. So that, for each arch, the vector of covariates is chosen to be $\mathbf{x}_{ij} = (1, S_i T_j, d_{ij})$.

4 Posterior Inference

We fix the hyperparameters in (3) - (4) to be $\mathbf{m} = (7, 0, 0)$, $\Sigma = \text{diag}(2, 1.5, 1.5)$, $\boldsymbol{\alpha} = (2, 2, 2, 2)$. The number of components K was selected among several possible values as the one giving clearer interpretability of the posterior cluster estimate; with $K = 4$ we were able to interpret 4 different behaviours in terms of geolocation of the source and target nodes of the edges.

Posterior inference was computed using Stan software [2]; MCMC chains were ran each for 2000 iterations after 2000 iterations of burn-in. Convergence was checked using both visual inspection of the chains and standard diagnostics available in the CODA package. From the comparison between the observed flow on the network (Figure 1 (b)) and the posterior expected value (Figure 1(c)), we see that our model correctly assigns zero flow to several edges, while predicting accurately the flow on the most travelled ones. It is clear that our model predicts quite well the observed flow.

Figure 2 reports a point estimate of the clustering structure of the edges arising from the Poisson mixture in (1), i.e. through the allocation variables identifying the mixture components. The point estimate has been obtained by finding the MCMC draw that minimizes the Binder loss function [1]. The first cluster includes all arcs with low flow and consists mainly of arcs belonging to the periphery of the city. The second group contains many arcs which are closer to the city center. The third cluster clearly groups together arcs with high flow in the middle of the city. Finally the last one includes only few arcs that represent the most travelled routes. The parameter β_{k3} , that represents the impact of the distance between node i and node j on the flow, is strongly negative for all the clusters, as expected, except the first one, where its values are positive although close to 0.

In all the cluster maps, the central big dot is Milan city center (Duomo), the other recognizable nodes are Cadorna train station (on the left of the map) and Centrale train station (top right), which, as expected, are focal points for the bike sharing system.

5 Competitor Models

Finally, we compare the predictive performances of our approach to the ones obtained by alternative models. Specifically, we consider model (1) without covariates (0infl), a standard Poisson mixture regression (i.e. model (1) - (2) with $\theta \equiv 0$) (Reg) and our model (Reg0infl). The corresponding priors are matched to give same a priori information on the same parameters. We report the comparison in

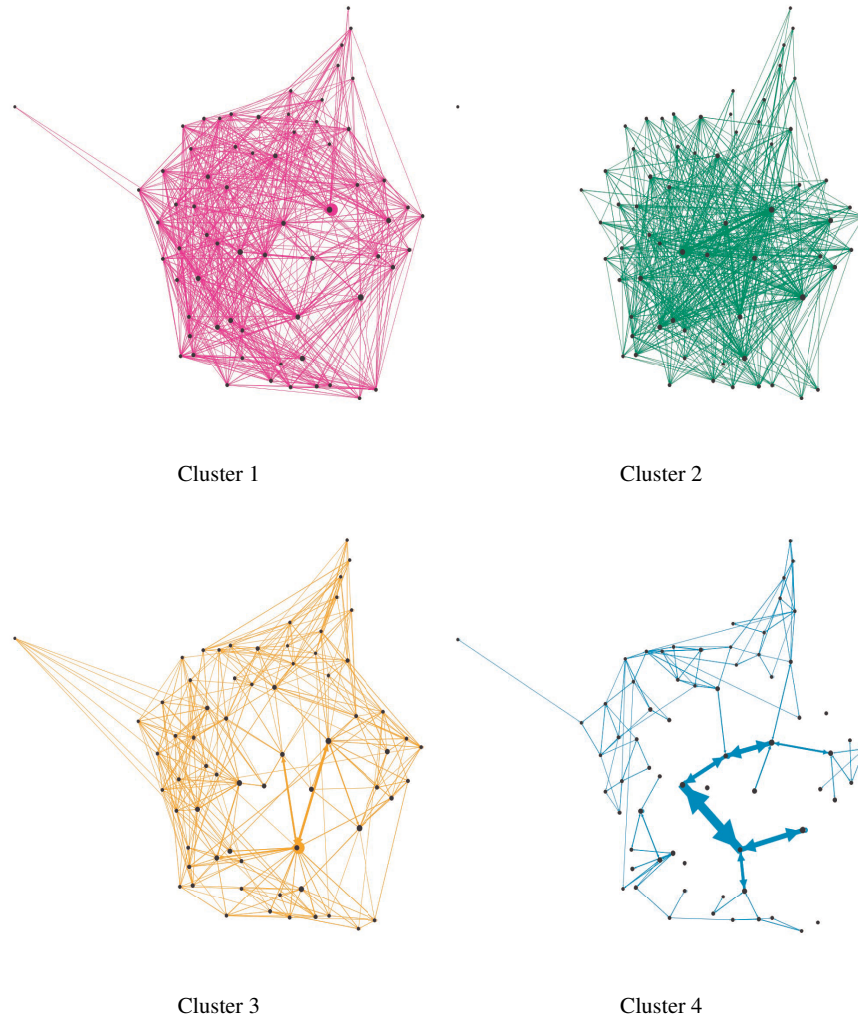


Fig. 2: The clustering of the edges arising from the Poisson mixture model.

Table 1, where for each model the following indexes of performance are shown:
 (i) *MSE*: the mean square error between the posterior mean and the observed flow.
 (ii) *LOO-ELPD*: the leave one out estimate of the expected log predictive density (computed using the `loo` package). (iii) *LPML*: the log pseudo marginal likelihood.

`Reg` and `Reg0infl` clearly outperform `0infl`, moreover, `Reg0infl` is slightly better than `Reg`. From a visual inspection of the predicted flows on both models, we can conclude that the main difference between `Reg0infl`'s and `Reg`'s prediction is that `Reg0infl` correctly assigns zero flow on a significant number of edges while `Reg` assigns to the same edges a small but positive flow.

Model	MSE	LOO-ELPD	LMPL
Reg0infl	1,518.3	-14,270.5	-11,237.9
Reg	1,685.4	-18,347.7	-18,341.7
0infl	300,373.4	-84,862.0	-103,089.3

Table 1: Predictive performances comparison.

6 Discussion and Conclusions

In this paper we have presented a full Bayesian model to analyze the mobility of one of the bike sharing systems in Milan. The proposed approach is based on modelling the counts of the number of travels between pairs of (clusters of) bike stations as the flow data on the edges of a complex network. We have assumed a zero-inflated Poisson mixture regression model to capture both the topology of the network and various range of behaviours in the flows, incorporating in the model also information coming from covariates such as the pairwise geographical distance between nodes.

Through MCMC simulations, we have shown how our model compares favorably against possible Bayesian competitors and how it describes the overall structure of the data. Nonetheless, we believe that incorporating into the model other information such as the proximity of a bike station to an underground stop or other places of interest might improve the overall quality of the prediction.

In the future, we aim at applying our model to the whole network and compare the inference. As an alternative we could also develop ad-hoc clustering strategies to reduce the dimensionality of the dataset, aggregating more nodes in the periphery of the city while keeping distinct the ones in the center.

Acknowledgement

We thank Clear Channel for having shared these data with us.

References

1. Binder, D.A.: Bayesian cluster analysis. *Biometrika* **65**(1), 31–38 (1978)
2. Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. *Journal of Statistical Software, Articles* **76**(1), 1–32 (2017). DOI 10.18637/jss.v076.i01. URL <https://www.jstatsoft.org/v076/i01>
3. Congdon, P.: A Bayesian approach to prediction using the gravity model, with an application to patient flow modeling. *Geographical Analysis* **32**(3), 205–224 (2000)
4. Ester, M., Kriegel, H.P., Xu, X.: Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. In: *International Symposium on Spatial Databases*, pp. 67–82. Springer (1995)
5. Frühwirth-Schnatter, S.: *Finite mixture and Markov switching models*. Springer Science & Business Media (2006)
6. Goldenberg, A., Zheng, A.X., Fienberg, S.E., Airoldi, E.M., et al.: A survey of statistical network models. *Foundations and Trends® in Machine Learning* **2**(2), 129–233 (2010)
7. Torti, A., Pini, A., Vantini, S.: Modeling time-varying mobility flows using function-on-function regression: analysis of a bike sharing system in the city of Milan. Tech. rep., MOX, Politecnico di Milano (2019)

Statistical classics in the big data era. When (astro-physical) models are nonregular

Statistica classica nell'era dei big data. Verosimiglianza e modelli non regolari

Alessandra R. Brazzale and Valentina Mameli

Abstract The hundreds of thousands of γ -ray counts collected by the *Fermi* Large Area Telescope (LAT) contribute uniquely to the study of the most extreme phenomena in our Universe such as active galactic nuclei, supernova remnants and pulsar wind nebula. Likelihood is currently being used to analyse these extensive data sets. The aim of this contribution is to raise awareness among the readers about the possible non-regularities inherent the corresponding parametric models and about the failures of the underlying likelihood-based theory. Especially practitioners may be less familiar with the resulting limiting distributions.

Abstract Un contributo fondamentale allo studio degli eventi astrofisici estremi del nostro Universo è stato dato dalle centinaia di migliaia di dati raccolti dal Telescopio Fermi. La teoria della verosimiglianza viene attualmente utilizzata per analizzare questi insiemi di dati di grande dimensione. Questo lavoro vuole sensibilizzare i lettori sulle possibili non regolarità inerenti i modelli parametrici applicati a questi studi e il conseguente fallimento della teoria basata sulla verosimiglianza. In particolare, potrebbe non essere loro nota la corrispondente teoria asintotica di riferimento.

Key words: Asymptotic inference, Boundary problem, Likelihood, Mixture distribution, Regularity condition

1 Introduction and motivation

Resolving the γ -ray sky by detecting as yet unidentified sources and accurately measuring the diffuse background emission is a declared key scientific objective

Alessandra R. Brazzale

Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy, e-mail: alessandra.brazzale@unipd.it

Valentina Mameli,

Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Via Torino 155, 30172 Venezia Mestre, Italy e-mail: valentina.mameli@unive.it

of the *Fermi* Gamma-ray Space Telescope collaboration.¹ Frequentist or Bayesian likelihood-based inference is the predominant point-source analysis method of astrophysics. The application of likelihood to photon-counting experiments was first described by [2]. [5, 6] applied it to COS-B γ -ray data and to X-ray data from the *Einstein Observatory*. Likelihood was furthermore used with data from the Energetic Gamma Ray Experiment Telescope² and to analyse data from the Imaging Compton Telescope³.

In astrophysics, the likelihood ratio statistic is in particular used to determine the statistical significance of a putative source. It is commonly believed that under the null hypothesis the three classical tests of likelihood based inference, that is, those based on the Wald, score and likelihood ratio statistics, are asymptotically equivalent and, to the first order of approximation, follow a chi-squared distribution. However, in order to hold true, this statement requires a number of regularity conditions. These may be formulated in several ways and require, among others, differentiability of the underlying joint probability or density function up to a suitable order and finiteness of the Fisher information matrix.

The aim of this contribution is to highlight a number of non-regularities inherent some of the parametric models which are used to analyse the γ -ray count maps provided by the Large Area Telescope (LAT) on board the *Fermi* spacecraft, and to raise awareness about how classical likelihood asymptotic theory may then fail. Section 2 presents the statistical classics we will focus upon. Section 3 introduces the reader to single-source and variable-source-number models. Section 4 presents the regularity conditions which are required for likelihood asymptotics to hold. The paper closes with Section 5 where we give some preliminary considerations about how to tackle possible nonregular situations. The discussion will be limited to so-called first order, or large-sample, asymptotics. Our aim however is to extend this research, which is still in its beginnings, to higher order, or small-sample, asymptotics.

2 Statistical classics

Consider a parametric statistical model with probability density or mass function $f(y; \theta)$, where the parameter θ takes values in a subset $\Theta \subseteq \mathbb{R}^p$, $p \geq 1$, and $y = (y_1, \dots, y_n)$ are n observations from $Y = (Y_1, \dots, Y_n)$. Let $L(\theta) = L(\theta; y) \propto f(y; \theta)$ and $l(\theta) = \log L(\theta)$ denote the likelihood and the log-likelihood functions, respectively. The maximum likelihood estimate (MLE) $\hat{\theta}$ of θ is the value of θ which maximises $L(\theta)$ or equivalently $l(\theta)$. Under mild regularity conditions on the log-likelihood function, which will be formulated in Section 4, $\hat{\theta}$ solves the score equation $u(\theta) = 0$, where $u(\theta) = \partial l(\theta) / \partial \theta$ is the score function. We furthermore define the observed information function $j(\theta) = -\partial^2 l(\theta) / \partial \theta \partial \theta^\top$ and the expected or Fisher information $i(\theta) = E[j(\theta; Y)]$, where θ^\top denotes transposition of θ .

¹ <https://fermi.gsfc.nasa.gov/>

² <https://heasarc.gsfc.nasa.gov/docs/cgro/egret/>

³ <https://heasarc.gsfc.nasa.gov/docs/cgro/comptel/>

The three classical likelihood based statistics for testing $\theta = \theta_0$ are the

$$\begin{aligned} \text{standardized MLE,} & \quad (\hat{\theta} - \theta_0)^\top j(\hat{\theta})(\hat{\theta} - \theta_0), \\ \text{score statistic,} & \quad u(\theta_0)^\top j(\hat{\theta})^{-1} u(\theta_0), \\ \text{likelihood ratio} & \quad 2\{l(\hat{\theta}) - l(\theta_0)\}, \end{aligned}$$

where the observed information $j(\hat{\theta})$ is at times replaced by the Fisher information $i(\theta)$. These statistics are also known under the names of Wald, Rao and Wilks tests, respectively. If the parametric model is regular, the finite sample null distribution of the above three statistics converges to a χ_p^2 distribution to the order $O(n^{-1})$ as $n \rightarrow \infty$. For θ scalar, inference may be based on the corresponding signed versions, that is, on the signed Wald statistic, $(\hat{\theta} - \theta_0)j(\hat{\theta})^{1/2}$, score statistic, $u(\theta_0)j(\theta_0)^{-1/2}$, and likelihood root,

$$r(\theta_0) = \text{sign}(\hat{\theta} - \theta_0)[2\{l(\hat{\theta}) - l(\theta_0)\}]^{1/2},$$

whose finite sample distributions converge to the standard normal distribution to the order $O(n^{-1/2})$.

Suppose now that the parameter $\theta = (\psi, \lambda)$ is partitioned into a p_0 -dimensional parameter of interest, ψ , and a vector of nuisance parameters λ of dimension $p - p_0$. Large-sample inference for ψ is commonly based on the profile log-likelihood function $l_p(\psi) = l(\psi, \hat{\lambda}_\psi)$, which we obtain by replacing the nuisance parameter λ in $l(\psi, \lambda)$ by the corresponding constrained maximum likelihood estimate $\hat{\lambda}_\psi$, that is, the value of λ which maximises the log-likelihood $l(\psi, \lambda)$ for fixed ψ . We may then define the profile Wald, score and likelihood ratio statistics for testing $\psi = \psi_0$ as for the scalar parameter case, but now in terms of the profile log-likelihood $l_p(\psi)$, with $u_p(\psi) = \partial l_p(\psi) / \partial \psi$ and $j_p(\psi) = \partial^2 l_p(\psi) / \partial \psi \partial \psi^\top$ being the profile score and profile observed information functions. The asymptotic null distribution of these statistics is a $\chi_{p_0}^2$ distribution up to the order $O(n^{-1})$. If ψ is scalar, the distributions of the corresponding signed versions, $(\hat{\psi} - \psi_0)j_p(\hat{\psi})^{1/2}$, $u_p(\psi_0)j_p(\psi_0)^{-1/2}$, and

$$r_p(\psi_0) = \text{sign}(\hat{\psi} - \psi_0)[2\{l_p(\hat{\psi}) - l_p(\psi_0)\}]^{1/2}, \quad (1)$$

may be approximated by standard normal distributions up to the order $O(n^{-1/2})$.

3 Likelihood analysis of astrophysical data

The γ -ray data from the *Fermi* LAT take the form of whole sky maps of γ -ray counts which are spatially projected onto rectangles indexed by Galactic coordinates. Traditional analyses are based on so-called *single-source* models which assess the presence of a possible new source on a pixel-by-pixel basis using significance testing; see Section 7.4 of [3]. The Poisson distribution is used to model the probability

$$p_i = \frac{\theta_i^{n_i} e^{-\theta_i}}{n_i!}$$

of observing n_i counts stemming from direction (x_i, y_i) , where $x_i \in (-180, 180)$ and $y_i \in (-90, 90)$ are the Galactic longitude and latitude, respectively, given that the mean number of counts predicted by the model is $\theta_i > 0$. The likelihood $L(\theta) = \prod_i p_i$, for $\theta = (\theta_1, \dots, \theta_N)$, is the product of the probabilities p_i for the N observed pixels (x_i, y_i) . The mean components of the model, θ_i , depends on an “active” source which emits from direction $(\mu_0^x, \mu_0^y) \in (-180, 180) \times (-90, 90)$, whose presence we want to test, an unknown number $n_0 > 0$ of photons to be estimated. It furthermore depends on a finite number of “inactive” sources with given photon counts n_k stemming from a finite number of known directions (μ_k^x, μ_k^y) , $k = 1, 2, \dots$. Thus, the statistical model for pixel (x_i, y_i) is characterised by

$$\theta_i = c_g B_g(x_i, y_i) + c_e B_e(x_i, y_i) + n_0 f(x_i, y_i; \mu_0^x, \mu_0^y) + \sum_k n_k f(x_i, y_i; \mu_k^x, \mu_k^y).$$

Here, $f(x_i, y_i; \mu_k^x, \mu_k^y)$ is a suitable density function which models the contribution of the photon observed at pixel (x_i, y_i) assuming that it was emitted from a source located at (μ_k^x, μ_k^y) , while $f(x_i, y_i; \mu_0^x, \mu_0^y)$ does the same but with respect to the putative source at (μ_0^x, μ_0^y) . The quantities $B_g(x_i, y_i)$ and $B_e(x_i, y_i)$ predict the number of photons that are emitted from the Galactic and extra-galactic diffuse background radiation. The parameters to be estimated are the unknown background intensities $c_g > 0$ and $c_e > 0$ and the “active” source count n_0 . This is commonly done by maximum likelihood. Note that estimates of each parameter are required to be non negative as negative values are unphysical.

Variable-source-number models address the problem from a more global perspective, as they simultaneously estimate the number of sources in the whole map without the need to separate the latter into smaller cells and to work on single pixels; see Section 7.3 of [3]. A recent example of application to the *Fermi* LAT data is [7]. In practice, we have no information whether the photon was emitted from a source or belongs to the background, nor do we know the number of emitting sources and their directions in space. This situation is well represented by a finite mixture model and translates into the marginal model

$$h(x_i, y_i; \mu^x, \mu^y, \sigma_b, \omega) = \omega_0 g_b(x_i, y_i; \sigma_b) + \sum_k \omega_k f(x_i, y_i; \mu_k^x, \mu_k^y),$$

where $g_b(\cdot; \sigma_b)$ represents the distribution of photons from the background which is indexed by an unknown scale parameter σ_b , while the $f(\cdot; \cdot)$ again model the signal of specific sources located at (μ_k^x, μ_k^y) , $k = 1, \dots, K$. The vector $\omega = (\omega_0, \omega_1, \dots, \omega_K)$ contains the mixing proportions ω_k which can be viewed as the intensity ω_0 of the background and of each source, that is, ω_k . This model is hence characterised by a set $\theta = \{\mu^x, \mu^y, \sigma_b, \omega\}$ of $3K + 2$ parameters, where μ^x and μ^y are now vector parameters which contain the unknown longitudes and latitudes of the K emitting sources. [7] furthermore assume that the number of undetected sources K is itself unknown and needs be estimated.

4 Nonregular models

Assessing that no point source exists at direction (μ_0^x, μ_0^y) in the single-source model amounts to testing whether $c_0 = 0$. From Wilks's theorem, the likelihood ratio is expected to be asymptotically distributed as χ_1^2 under the null. Monte Carlo simulation carried out in [4] showed that the distribution is actually a $\tilde{\chi}_1^2$, that is, mixture of a point mass in 0 and a χ_1^2 . This is typical of a *boundary problem*.

Boundary problems represent the first and most extensively explored nonregular setting. A boundary problem arises when the value θ_0 specified by the null hypothesis, or parts of it, fall on the boundary of the parameter space. Informally, the statistical issues in likelihood-based inference occur because the maximum likelihood estimate can only fall on the side of θ_0 that belongs to the parameter space Θ . This implies that if the maximum occurs on the boundary, the score function need not be zero and the distributions of the related likelihood statistics will not converge to the typical normal or chi-squared distributions.

Here, we will assume that the following five conditions on the model function $f(y; \theta)$ hold.

Condition 1 All components of θ are identifiable. That is, two model functions $f(y; \theta^1)$ and $f(y; \theta^2)$ defined by any two different values $\theta^1 \neq \theta^2$ of θ are distinct almost surely.

Condition 2 The support of $f(y; \theta)$ does not depend on θ .

Condition 3 The parameter space Θ is a compact subset of \mathbb{R}^p , for a fixed positive integer p , and the true value θ^0 of θ is an interior point of Θ .

Condition 4 The partial derivatives of the log-likelihood function $l(\theta; y)$ with respect to θ up to the order three exist in a neighbourhood of the true parameter value θ^0 almost surely. Furthermore, in such a neighbourhood, n^{-1} times the absolute value of the log-likelihood derivatives of order three are bounded above by a function of Y whose expectation is finite.

Condition 5 The first two Bartlett identities hold, which imply that

$$E[u(\theta; Y)] = 0, \quad i(\theta) = \text{Var}[u(\theta; Y)].$$

Models which satisfy these requirements are said to be “regular” and cover a wide range of applications. However, there are many important cases where one or more conditions break down. These situations are not mere mathematical artifacts, but include many models of practical interest, such as the mixture distributions used in variable-source-number models, and change-point problems.

5 Concluding remarks

In the absence of a unifying theory, most of the individual problems have been treated on their own. The comprehensive review in [1] groups them into three broad

classes. The first considers the case where the parameter space is bounded and embraces, in particular, testing for a value of the parameter which lies on its boundary. A second class of problems concerns models where one part of the parameter vanishes when the remaining one is set to a particular value. The best-studied example of indeterminate parameter problem is the case of finite mixture models. Change-point problems are the third broad class of nonregular models. Most articles investigate the consequences of the failure of one regularity condition at a time. Mixture distributions and change-point problems deserve special attention as they represent situations where two conditions fail simultaneously.

Many problems can be dealt with rather straightforwardly. The limiting distribution of the likelihood ratio for testing for a boundary value of the parameter can generally be traced back to a χ^2 distribution with a suitable number of components and mixing weights which depend on the number of parameters on the boundary and on the design matrices in regression problems. This is also the only type of problem for which higher order asymptotic results are available. Other situations, such as testing for mixture components, require rather sophisticated tools such as limit theorems and extreme value theory for stationary and non-stationary random fields. In these cases, simulation-based approaches are often preferred to obtain the required significance levels.

Acknowledgements This project was supported by SID 2018 grant “Advanced statistical modelling for indexing celestial objects” (BIRD185983) awarded by the Department of Statistical Sciences of the University of Padova. Special thanks goes to Ruggero Bellio, Anthony C. Davison and Nancy Reid for most useful discussion on the subject.

References

1. Brazzale, A.R. and Mameli, V.: Likelihood asymptotics in nonregular settings: A review with emphasis on the likelihood ratio. *Working Paper Series* **18/4**, Department of Statistical Sciences, University of Padova (2018).
2. Cash, W.: Parameter estimation in astronomy through application of the likelihood ratio. *The Astrophysical Journal* **228**, 939–947 (1979).
3. Hobson, M.P., Rocha, G. and Savage, R.S.: Bayesian source extraction. In “Bayesian Methods in Cosmology” (Hobson, M.P. et al. Eds.), Cambridge University Press (2010).
4. Mattox, J.R., Bertsch, D.L. et al.: The likelihood analysis of Egret data. *The Astrophysical Journal* **461**, 396 (1996).
5. Pollock, A.M.T., Bignami, G.F. et al.: Search for gamma-radiation from extragalactic objects using a likelihood method. *Astronomy and Astrophysics*, **94**, 116 (1981).
6. Pollock, A.M.T.: The Einstein view of the Wolf-Rayet stars. *The Astrophysical Journal*, **320**, 283 (1987).
7. Sottosanti, A., Costantin, D., Bastieri, D. and Brazzale, A.R.: Discovering and locating high-energy extra-galactic sources by Bayesian mixture modelling. To appear in Springer Proceedings in Mathematics & Statistics – SIS 2017 “New Statistical Developments in Data Science” (A. Petrucci et al. Eds.), Springer-Verlag (2019).

Bayesian Variable Selection for High Dimensional Logistic Regression

Selezione bayesiana delle variabili nel modello di regressione logistica ad alta dimensionalità

Claudio Busatto, Andrea Sottosanti and Mauro Bernardi

Abstract This paper introduces a novel Bayesian approach to the problem of variable selection in high-dimensional logistic regression. In particular, we present a Marginalized Reversible Jump MCMC (MRJ) algorithm and its extensions, that exploits the data-augmentation structure using the Pólya–Gamma distribution. The proposed methods have been tested on simulated datasets, showing good performances in selecting the relevant regressors.

Abstract *In questo articolo presentiamo un nuovo approccio bayesiano al problema della selezione delle variabili nel modello di regressione logistica ad alta dimensionalità. In particolare, vengono introdotti l'algoritmo Marginalized Reversible Jump (MRJ) e le sue generalizzazioni, che si basano sulla rappresentazione aumentata con l'introduzione della variabile Pólya–Gamma. Questi metodi sono stati valutati su dataset simulati, mostrando una buona capacità di selezionare i regressori rilevanti.*

Key words: logistic regression, Bayesian variable selection, high-dimensional regression, Pólya–Gamma, reversible-jump, multiple-try.

Claudio Busatto

University of Padova, Department of Statistical Sciences, Via Cesare Battisti, 241, 35121 Padova, Italy, e-mail: claudio.busatto@studenti.unipd.it

Andrea Sottosanti

University of Padova, Department of Statistical Sciences, Via Cesare Battisti, 241, 35121 Padova, Italy, e-mail: sottosanti@stat.unipd.it

Mauro Bernardi

University of Padova, Department of Statistical Sciences, Via Cesare Battisti, 241, 35121 Padova, Italy, e-mail: mauro.bernardi@unipd.it

1 Introduction

Genetic datasets are usually characterised by a large number of covariates if compared to the number of available observations. This property denies the use of classical methods to analyse such data and justifies the application of Bayesian techniques.

This paper deals with variable selection in high-dimensional logistic regression. Specifically, we rely on the Pólya–Gamma augmentation approach of [4], that allows to analytically integrate the regression parameters out of the joint posterior distribution, as for the linear regression model (see, e.g., [2]). However, when the number of covariates is large, complete model enumeration is not possible, and we rely on reversible jump type algorithms that efficiently explore the space of competing models towards optimal solutions. The reversible jump idea is then combined with the data augmentation approach to provide a marginalised reversible jump algorithm (MRJ) for binary data. Leveraging the Multiple–Try Metropolis–Hastings (MTMH) idea we further extend the RJ approach (see, e.g., [1]) to simultaneously dealing with multiple proposals.

2 Bayesian variable selection

2.1 The statistical model

Let $\mathbf{y} = (y_1, \dots, y_n)$ be a binary response vector from a $Y_i \sim \text{Be}(\psi_i)$ distribution, where $\psi_i = 1/(1 + e^{-(\alpha + \mathbf{x}_i \beta)})$ and $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n$, is the set of p covariates related to the i -th observation. The model includes a constant $\alpha \in \mathbb{R}$, while $\beta = (\beta_1, \dots, \beta_p)^\top$ is the p -dimensional vector of regression parameters.

Following the data-augmentation scheme in [4], it can be shown that, by augmenting the observed data with a Pólya–Gamma latent variable, $\pi(\omega_i) \sim \text{PG}(1, 0)$, $i = 1, \dots, n$, the multivariate Normal prior distribution on the regression parameters $\vartheta = (\alpha, \beta)^\top$ is conjugated with the likelihood function $\phi_n(\mathbf{X}\vartheta, \mathbf{W}^{-1})$. Therefore, assuming $\pi(\vartheta) \sim \text{N}_p(\mu, \sigma^2 \Sigma)$, the following set of full-conditional distributions are analytically available

$$\begin{aligned} \pi(\omega | \mathbf{y}, \mathbf{X}, \vartheta) &\sim \text{PG}(1, \mathbf{X}\vartheta) \\ \pi(\vartheta | \mathbf{y}, \mathbf{X}, \vartheta) &\sim \text{N}_p(\Sigma^*(\mathbf{X}'\mathbf{W}\tilde{\mathbf{y}} + (\sigma^2 \Sigma)^{-1}\mu), \Sigma^*), \end{aligned} \quad (1)$$

where $\Sigma^* = (\mathbf{X}'\mathbf{W}\mathbf{X} + (\sigma^2 \Sigma)^{-1})^{-1}$, $\tilde{\mathbf{y}} = \left(\frac{y_1 - 0.5}{\omega_1}, \dots, \frac{y_n - 0.5}{\omega_n}\right)$ and $\mathbf{W} = \text{diag}(\omega_1, \dots, \omega_n)$. In order to control the posterior variability of ϑ , the prior variance–covariance matrix is allowed to depend on an hyperparameter σ^2 , which is assumed to be a realisation of an Inverse–Gamma distribution, $\pi(\sigma^2) \sim \text{IG}(\lambda, \nu)$.

Hereafter, regression models are indexed by a p -dimensional vector $\gamma = (\gamma_1, \dots, \gamma_p)$, where $\gamma_j = 1$, $j = 1, \dots, p$, when the j -th regressor is included in the model, 0 other-

wise, and the complexity of the model is $p_\gamma = \sum_{j=1}^p \gamma_j$. In particular, the parameter γ follows a Binomial distribution, $\pi(\gamma) \sim \text{Bin}(p, \phi)$, where the hyperparameter ϕ controls for the prior probability of inclusion of each covariate. Under previous assumptions, the joint posterior distribution is

$$\begin{aligned} \pi(\vartheta, \gamma, \sigma^2 \mid \mathbf{y}, \mathbf{X}, \omega) &\propto \pi(\vartheta \mid \gamma, \sigma^2) \pi(\gamma) \pi(\sigma^2) \prod_{i=1}^n \mathbb{P}(Y_i = y_i \mid \mathbf{x}_i, \omega_i, \vartheta) \pi(\omega_i) \\ &\propto \exp \left\{ -\frac{1}{2} (\tilde{\mathbf{y}} - \mathbf{X}_\gamma \vartheta_\gamma)' \mathbf{W} (\tilde{\mathbf{y}} - \mathbf{X}_\gamma \vartheta_\gamma) \right\} \\ &\quad \times \pi(\vartheta \mid \gamma, \sigma^2) \pi(\gamma) \pi(\sigma^2), \end{aligned} \quad (2)$$

where $\pi(\vartheta \mid \gamma, \sigma^2) \sim \mathcal{N}_{p_\gamma}(\mu_\gamma, \sigma^2 \Sigma_\gamma)$ and $\mathbf{X}_\gamma \in \mathbb{R}^{n \times p_\gamma}$ denotes the data matrix containing the j -th column only if $\gamma_j = 1$, $j = 1, \dots, p$. Under the previous prior structure, the set of full-conditional distributions of $\omega, \vartheta, \sigma^2$ can be updated as

$$\begin{aligned} \pi(\omega \mid \mathbf{X}, \gamma, \vartheta) &\sim \text{PG}(1, \mathbf{X}_\gamma \vartheta_\gamma), \\ \pi(\vartheta \mid \mathbf{y}, \mathbf{X}, \omega, \gamma, \sigma^2) &\sim \mathcal{N}(\Sigma_\gamma^* (\mathbf{X}_\gamma' \mathbf{W} \tilde{\mathbf{y}} + (\sigma^2 \Sigma_\gamma)^{-1} \mu_\gamma), \Sigma_\gamma^*), \\ \pi(\sigma^2 \mid \vartheta, \gamma) &\sim \text{IG} \left(\lambda + \frac{p_\gamma}{2}, \nu + \frac{1}{2} \sum_{j=1}^{p_\gamma} \vartheta_j^2 \right), \end{aligned} \quad (3)$$

where $\Sigma_\gamma^* = (\mathbf{X}_\gamma' \mathbf{W} \mathbf{X}_\gamma + (\sigma^2 \Sigma_\gamma)^{-1})^{-1}$. Integrating out ϑ from (2) yields

$$\begin{aligned} \pi(\gamma, \sigma^2 \mid \mathbf{y}, \mathbf{X}, \omega) &\propto \pi(\gamma) \pi(\sigma^2) |\mathbf{X}_\gamma' \mathbf{W} \mathbf{X}_\gamma + (\sigma^2 \Sigma_\gamma)^{-1}|^{-1/2} |\sigma^2 \Sigma_\gamma|^{-1/2} \\ &\quad \times \exp \left\{ \frac{1}{2} \tilde{\mathbf{y}}' \mathbf{W} \mathbf{X}_\gamma (\mathbf{X}_\gamma' \mathbf{W} \mathbf{X}_\gamma + (\sigma^2 \Sigma_\gamma)^{-1})^{-1} \mathbf{X}_\gamma' \mathbf{W} \tilde{\mathbf{y}} \right\}. \end{aligned} \quad (4)$$

At each iteration, parameter γ is updated with a Metropolis–Hastings step, with (4) as the target distribution. New values of the chain are sampled from the proposal distribution

$$q(\gamma' \mid \gamma) = \frac{1}{\binom{p}{d}}, \quad \text{if } \sum_{j=1}^p |\gamma'_j - \gamma_j| = d, \quad (5)$$

which is symmetric in γ' and γ . In this way, all models with $(p_\gamma - d)$ and $(p_\gamma + d)$ are taken into account with the same probability.

2.2 Simulation algorithms

The novel Generalized Marginalized Multiple-Try Reversible-Jump (GMMT-RJ) algorithm is described in 1. The prior distribution for ϑ is $\pi(\vartheta \mid \gamma, \sigma^2) \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_\gamma)$, simplifying the posterior covariance matrix to $\Sigma_\gamma^* = (\mathbf{X}_\gamma' \mathbf{W} \mathbf{X}_\gamma + \frac{1}{\sigma^2} \mathbf{I}_{p_\gamma})^{-1}$. We refer to the simplest case where $k = 1$ as Marginalized Reversible-Jump (MRJ) algo-

rithm. However, when the number of covariates p is extremely high, the simulated chain needs to be very long in order to explore the target distribution exhaustively. Therefore, the Multiple-Try MH (MMT-RJ) generalisation here proposed allows to evaluate more candidates at each iteration. To further increase the flexibility of the model, different proposal can be implemented (GMMT-RJ), see, e.g., [1].

In the proposed algorithms, the model indicators γ' are generated from (5) by sampling a d -dimensional indexing vector \mathbf{I} and setting the d components of γ corresponding to \mathbf{I} equal to 0, if $\gamma(\mathbf{I}) = 1$, and equal to 1, otherwise. Because of the symmetry of (5), the importance weights simplify to $w(\gamma' | \gamma) = \pi(\gamma', \sigma^2 | \mathbf{y}, \mathbf{X}_\gamma, \omega)$. As a result, the MH acceptance probability $\alpha(\gamma, \gamma')$ in Algorithm 1 is proven to satisfy the detailed balance condition, see, e.g., [1].

3 Numerical examples

Let $\gamma^* = (\gamma_1^*, \dots, \gamma_p^*)$ be the data generator which includes j -th variable if $\gamma_j^* = 1$, $j = 1, \dots, p$. The MMT-RJ algorithm has been tested on a simulated dataset containing $n = 200$ observations and $p = 300$ covariates. In our simulation settings, the true model γ^* has complexity $p_{\gamma^*} = 26$. The random observation y_i , $i = 1, \dots, n$, is sampled from a Bernoulli distribution, $\text{Be}(\pi_i)$, where $\pi_i = e^{\mathbf{x}_i \boldsymbol{\vartheta}} / (1 + e^{\mathbf{x}_i \boldsymbol{\vartheta}})$, $x_{ij} \sim \text{N}(0, 1) \forall i, j$ and $\boldsymbol{\vartheta}_j$, $j = 1, \dots, p$, are arbitrarily chosen by sampling from a $\text{U}[2, 5]$ distribution when $\gamma_j^* = 1$, and 0 otherwise. To visualise the results, we introduce the vector \mathbf{R} that enumerates the visited models, having the r -th element defined as $r^{(r)} = 1 + \gamma^{(r)} \mathbf{d}$, where $\gamma^{(r)}$ is the model visited at iteration r and $\mathbf{d} = [2^0 \ 2^1 \ \dots \ 2^{p-1}]^\top$, see, e.g., [5]. The parameters k and ϕ are treated as tuning parameters for the MH acceptance probability. Figure 1a shows the models visited at each iteration of the chain for $\phi = 0.3$ and $d = 1$. Interestingly, the post burn-in exploration of the target is limited to a narrow range of models, which includes the true model γ^* denoted by the red dashed line. Moreover, the post burn-in acceptance probability ratio of the chain is 61%. The posterior inclusion probability, π_j^* , $j = 1, \dots, p$, is given by

$$\pi_j^* = \sum_{r=b+1}^R \mathbb{1}(\gamma_j^{(r)} = 1) / (R - b - 1), \quad (6)$$

where b is the burn-in period. Variables are included as regressors if π_j^* is greater than an arbitrarily chosen threshold, $c \in (0, 1)$. The vector π^* and the true model γ^* allow the construction of a ROC curve to test the performance of the method, see Figure 1b. When $c = 0.5$, the sensitivity reaches 81%, while the specificity is 95%. The maximum sensitivity is 96% for $c = 0.44$. To evaluate the goodness of the variable selection, the *AUC* statistic is considered. This is equal to 0.94, suggesting that the proposed method provides an accurate selection of the relevant regressors.

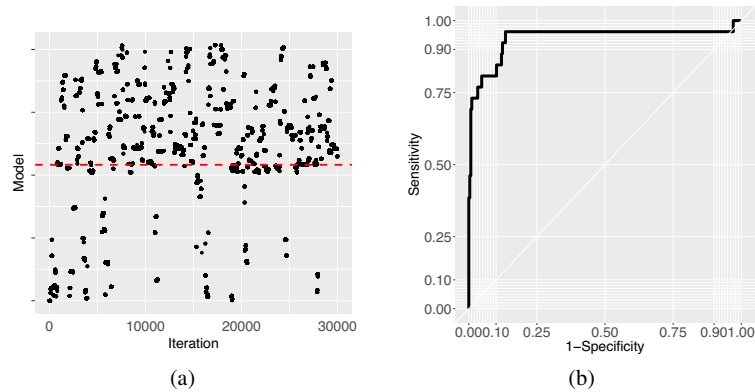


Fig. 1: Regression models visited by the chain (1a) and ROC curve (1b) for the MMT-RJ algorithm, with $\phi = 0.3$ and $d = 1$.

References

1. Casarin, R., Craiu, R. & Leisen, F. Stat Comput (2013) 23: 185. <https://doi.org/10.1007/s11222-011-9301-9>
2. George, E. I., & McCulloch, R. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, 7(2), 339-373.
3. John M. Chambers (1971) Regression Updating, *Journal of the American Statistical Association*, 66:336, 744-748, DOI: 10.1080/01621459.1971.10482338
4. Nicholas G. Polson, James G. Scott & Jesse Windle (2013): Bayesian Inference for Logistic Models Using PlyaGamma Latent Variables, *Journal of the American Statistical Association*, 108:504, 1339-1349, DOI: 10.1080/01621459.2013.829001
5. S.P Brooks, P Giudici & A Philippe (2003) Nonparametric Convergence Assessment for MCMC Model Selection, *Journal of Computational and Graphical Statistics*, 12:1, 1-22, DOI: 10.1198/1061860031347

Algorithm 1: Generalized Marginalized Multiple-Try Reversible Jump with latent variable Polya-Gamma

 Initialize parameters γ_0 , σ_0^2 , and ω_0 by sampling from their prior distribution;

for r from 1 to R **do**

 [1] **for** k from 1 to K **do**

 Let \mathbf{I} be an indexing vector and sample $\mathbf{I}_1, \dots, \mathbf{I}_k$ values from set $\{1, \dots, p\}$ with probability $1/p$;

 Set $\{\gamma_i^{(k)}\}_{i \in \mathbf{I}} = 1 - \{\gamma_i^{(r-1)}\}_{i \in \mathbf{I}}$ and compute the weights

$$w_k(\gamma^{(k)} | \gamma^{(r-1)}) = \pi(\gamma^{(k)}, \sigma^2 | \mathbf{y}, \mathbf{X}_{\gamma^{(k)}}, \omega);$$

end

 [2] Select $\gamma^{(j)}$ according to the probability density

$$\bar{w}_k = \frac{w_j(\gamma^{(j)} | \gamma^{(r-1)})}{\sum_{k=1}^K w_k(\gamma^{(k)} | \gamma^{(r-1)})};$$

 [3] Set $\mathbf{v}^{(j)} = \gamma^{(r-1)}$ and sample $\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(j-1)}, \mathbf{v}^{(j+1)}, \dots, \mathbf{v}^{(K)}$ auxiliary values:

for k from 1 to K and $k \neq j$ **do**

 Let \mathbf{I}' be an indexing vector and sample $\mathbf{I}'_1, \dots, \mathbf{I}'_k$ values from set $\{1, \dots, p\}$ with probability $1/p$;

 Set $\{\mathbf{v}_i^{(k)}\}_{i \in \mathbf{I}'} = 1 - \{\gamma_i^{(j)}\}_{i \in \mathbf{I}'}$ and compute the weights

$$w_k(\mathbf{v}^{(k)} | \gamma^{(j)}) = \pi(\mathbf{v}^{(k)}, \sigma^2 | \mathbf{y}, \mathbf{X}_{\mathbf{v}^{(k)}}, \omega);$$

end

[4] Compute

$$\alpha(\gamma^{(r-1)}, \gamma^{(j)}) = \min \left\{ 1, \frac{\sum_{k=1}^K w_k(\gamma^{(k)} | \gamma^{(r-1)})}{\sum_{k=1}^K w_k(\mathbf{v}^{(k)} | \gamma^{(j)})} \right\};$$

 [5] With probability α **do**:

 Set $\hat{\gamma}^{(r)} = \gamma^{(j)}$,

 Sample $\hat{\vartheta}^{(r)}$ from a $\mathcal{N}(\Sigma_{\gamma^{(r)}}^* (\mathbf{X}'_{\gamma^{(r)}} \mathbf{W} \tilde{\mathbf{y}}), \Sigma_{\gamma^{(r)}}^*)$,

 Sample $\hat{\sigma}_{(r)}^2$ from a $\text{IG} \left(\lambda + \frac{p_{\gamma^{(r)}}}{2}, \mathbf{v} + \frac{1}{2} \sum_{j=1}^{p_{\gamma^{(r)}}} \vartheta_j^2 \right)$,

 Sample $\hat{\omega}^{(r)}$ from a $\text{PG}(1, \mathbf{X}_{\gamma^{(r)}} \vartheta_{\gamma^{(r)}}^{(r)})$, using `pgdraw` in the R package

`pgdraw`;

[6] Otherwise set:

$$\begin{aligned} \hat{\gamma}^{(r)} &= \gamma^{(r-1)}, & \hat{\vartheta}^{(r)} &= \vartheta^{(r-1)}, \\ \hat{\sigma}_{(r)}^2 &= \sigma_{(r-1)}^2, & \hat{\omega}^{(r)} &= \omega^{(r-1)}. \end{aligned}$$

end

Bayesian modeling for large spatio-temporal data: an application to mobile networks

Modelli bayesiani per grandi dataset spazio-temporali: un'applicazione a dati di telefonia mobile

Annalisa Cadonna, Andrea Cremaschi, Alessandra Guglielmi

Abstract Spatio-temporal areal data can be seen as a collection of time series which are spatially correlated, according to a specific neighboring structure. We propose a hierarchical Bayesian model for spatio-temporal areal data, which allows for spatial model-based clustering. Then, we develop efficient MCMC algorithms based on numerical linear algebra, which exploit the sparse structure of the precision matrix of a spatio-temporal Gaussian Random Markov Field (GMRF). Finally, we present an application to mobile data, with the goal to model, predict and spatially cluster population density dynamics.

Abstract I dati areali spazio-temporali possono essere interpretati come una collezione di serie temporali correlate spazialmente, secondo una specifica struttura di vicinato. Qui proponiamo un modello gerarchico bayesiano per dati areali spazio-temporali, che consente di fare clustering “model-based”. Inoltre, sviluppiamo algoritmi efficienti di tipo MCMC, basati su risultati di algebra lineare numerica, sfruttando la sparsità della matrice di precisione di un Gaussian Markov Random Field che dipende dall'area e dal tempo. Studiamo un'applicazione a dati di telefonia mobile, con l'obiettivo di modellizzare, predire e raggruppare spazialmente dinamiche di densità di popolazione.

Key words: Spatio-temporal models, clustering of time series, Bayesian semiparametric models

Annalisa Cadonna

WU Vienna University for Economics and Business e-mail: annalisa.cadonna@wu.ac.at

Andrea Cremaschi

University of Oslo e-mail: andrea.cremaschi@medisin.uio.no

Alessandra Guglielmi

Dipartimento di Matematica, Politecnico di Milano, e-mail: alessandra.guglielmi@polimi.it

1 Introduction

In this paper we introduce a hierarchical Bayesian model for spatio-temporal areal data, which allows for spatial model-based clustering. We want to study population density dynamics in the municipality of Milano, using mobile phone data. The data under investigation is a large collection of long time series which are spatially correlated, according to a specific neighboring structure. We want also to cluster the metropolitan area of Milano in subregions sharing similar population dynamic pattern along time.

A few Bayesian approaches have been proposed to deal with both spatial and temporal dependence on areal unit data as, for instance, in [5], where the authors introduce the temporal dependence through an autoregressive structure on the means of a Gaussian Markov Random Field (GMRF) over time. The model in [5] allows for clustering through a parametric mixture on the spatio-temporal parameters. Our model can be seen as an extension of [5]. The main modeling difference is that we introduce areal-specific autoregressive parameters and use a nonparametric prior for model-based clustering. From a computational point of view, we exploit the fact our spatio-temporal model can be seen as a spatio-temporal GMRF and combine the algorithms in [4] and [8] to propose a MCMC algorithm for efficient posterior simulation. We would like to point out that an alternative to doing inference based on MCMC simulation is to use a Integrated Nested Laplace Approximations (INLA, [9]). INLA has become increasingly popular to fit spatio-temporal modeling, also thanks to the R-INLA package. However, as it will be clear from the next sections, while R-INLA can fit many different spatio-temporal models, we cannot fit our model with it.

This paper brings several contributions to the analysis of spatio-temporal areal datasets. First, we propose a hierarchical Bayesian model for spatio-temporal areal data, which allows for spatial model-based clustering. Moreover, the Bayesian approach allows to deal naturally with missing values. Second, we develop efficient algorithms based on numerical linear algebra, which exploit the sparse structures of the precision matrix of a spatio-temporal GMRF. Our algorithms are implemented in R. Due to the restriction on the number of pages we omit details on the MCMC algorithm. Finally, we present an application to mobile data, with the goal to model, predict and spatially cluster population density dynamics.

2 Description of the dataset

The dataset we consider here is provided by Telecom Italia, the biggest mobile company in Italy, as part of the Green Move Initiative, through a research agreement between Telecom and Politecnico di Milano. The dataset has been previously described and analyzed in [7] and [10], over a finer spatial and temporal grid. For this work, the metropolitan area of Milano is partitioned into a grid of $I = 24 \times 27 = 648$ sites. The data are considered every two hours, from March 18th, 2009, until March

31st, 2009. This means that, for each of the 648 lattice sites, we have a time series of length $T = 156$, that is a total of 101,088 observations. The Erlang number is calculated as the sum of the length of all the calls in a given time interval divided by the length of the interval; equivalently, it can be seen as the average number of mobile phones simultaneously calling through the network. The Erlang can be considered proportional to the number of active users. When the Erlang number is recorded over all areal units in a region, it can be used as a proxy for population density in that region and, as it changes over times, of population density dynamics. Figure 1(a) shows the data recorded on Wednesday, March 18, 2009 at noon, for the entire metropolitan area. The financial district in the center of the city is identifiable, as it is the area with high mobile activities during working hours. We can also localize, for example, the Linate airport and the Monza area. Figure 1(b) shows the time series for 10 areal units selected at random. As we can see, the mobile activity is higher during the day hours and low at night. Moreover, we can see a difference between weekdays and weekends.

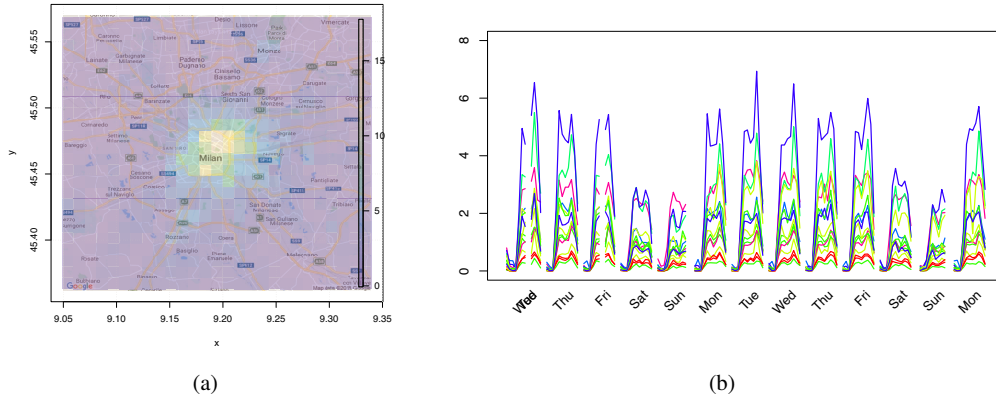


Fig. 1: **(a)**: Data recorded on Wednesday March 18, 2009, at noon. **(b)**: Time series in 10 areal units.

3 Model specification

Denote the observation for areal unit i ($i = 1, \dots, I$) and time t ($t = 1, \dots, T$) as Y_{it} . In our application, the response variable is the logarithm of the Erlang number. We model the observations through a Normal distribution, that is, for $i = 1, \dots, I$ and $t = 1, \dots, T$,

$$Y_{it} | \boldsymbol{\beta}_{i\phi_t}, w_{it}, \sigma_1^2 \stackrel{ind}{\sim} N(E[Y_{it} | \boldsymbol{\beta}_{i\phi_t}, w_{it}, \mathbf{h}_t], \sigma_1^2). \quad (1)$$

We also model the expected value in (1) through an harmonic regression equation, taking into account the weekend/weekday effect:

$$E[Y_{it} | \boldsymbol{\beta}_{i\phi_t}, w_{it}, \mathbf{h}_t] = \mathbf{h}_t \boldsymbol{\beta}_{i\phi_t} + w_{it}, \quad (2)$$

where \mathbf{h}_t is a $2p + 1$ dimensional vector of harmonic regressors, which adjusts for the seasonal temporal components in the observations. For each areal unit i , at each time, we have the same vector of harmonic regressors, that is $\mathbf{h}_t = (1, \cos(\omega_1 t), \sin(\omega_1 t), \dots, \cos(\omega_p t), \sin(\omega_p t))$. The vector $\boldsymbol{\beta}_{i\phi_t} = (\beta_{i\phi_t,1}, \dots, \beta_{i\phi_t,2p+1})'$ is the vector of area-specific coefficients, and w_{it} is the spatio-temporal random effect associated with areal unit i and time t . We assume for $\tilde{\mathbf{w}}_t = (w_{1t}, \dots, w_{It})'$, for $t = 2, \dots, T$, the prior

$$\tilde{\mathbf{w}}_t | \tilde{\mathbf{w}}_{t-1} \sim N_I(\text{diag}(\boldsymbol{\xi}) \tilde{\mathbf{w}}_{t-1}, \tau^2 Q(\rho, W)^{-1}), \quad (3)$$

where $\boldsymbol{\xi} = (\xi_1, \dots, \xi_I)$ is the vector of autoregressive coefficients, and $\text{diag}(\boldsymbol{\xi})$ the diagonal matrix with the regression coefficients on the diagonal. For the initial time $t = 1$, we assume $\tilde{\mathbf{w}}_1 \sim N_I(\mathbf{0}, \tau^2 Q(\rho, W)^{-1})$.

As far as modeling the spatial association, it is clear that the choice of the matrix $Q(\rho, W)$ in (3) is critical. W is the $I \times I$ matrix based on the contiguity structure of the I areal units. Specifically, $W_{i,j} = 1$ if areal unit i and areal unit j are neighbors. Here, we define as neighbors of a site i the 8-th cells surrounding i . We choose $Q(\rho, W) = \rho(\text{diag}(W\mathbf{1}) - W) + (1 - \rho)\mathbb{I}_I$, where \mathbb{I}_I is the $I \times I$ identity matrix and $\mathbf{1}$ is a $I \times 1$ vector of ones. This structure of the matrix $Q(\rho, P)$ was proposed in [6]. This model allows to estimate spatial correlation among the random effects. In fact, the parameter ρ controls the spatial autocorrelation structure: $\rho = 1$ corresponds to the intrinsic CAR prior ([1]), where the conditional expectation is the mean of the random effects in geographically adjacent areal units; on the other hand, $\rho = 0$ corresponds to independent random effects.

As mentioned in Section 1, we are interested in detecting which units share similar temporal patterns. For this reason, we allow for spatial clustering in our model through a Bayesian nonparametric prior on the coefficients $\boldsymbol{\beta}_{i\phi_t}$ and ξ_i , $i = 1, \dots, I$. In particular, we assume a ANOVA-DDP process prior ([3]) on $(\boldsymbol{\beta}_{i\phi_t}, \xi_i)$, i.e..

$$\begin{aligned} (\boldsymbol{\beta}_{i\phi_t}, \xi_i) | P_x &\stackrel{\text{ind}}{\sim} P_{\phi_t}, \quad i = 1, \dots, I, \\ \mathcal{P} = (P_{\phi_t}, \phi_t \in \{0, 1\}) &\sim \text{ANOVA-DDP}(\alpha, P_0), \end{aligned} \quad (4)$$

where $\text{ANOVA-DDP}(\alpha, P_0)$ represents the ANOVA Dirichlet process of mass parameter α and centering measure P_0 defined on $\mathbb{R}^{2(2p+1)+1}$. Marginally, each P_{ϕ_t} , $\phi_t \in \{0, 1\}$ follows a Dirichlet Process. We take the centering measure $P_0(d\boldsymbol{\beta}, d\xi)$ to be equal to a $N_{2(2p+1)+1}(d\boldsymbol{\beta} | \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \text{Beta}_{(-1,1)}(d\xi | \alpha_\xi, \beta_\xi)$. Here, $\text{Beta}_{(l,u)}(a, b)$ denotes a non standard Beta distribution with support on (l, u) . Since the DDP prior in (4) is a discrete prior, it can be equivalently represented through the introduction of a vector of allocation variables $\mathbf{s} = (s_1, \dots, s_I)$ and a set of unique values for the parameters $(\boldsymbol{\xi}^*, \boldsymbol{\beta}^*)$, such that $s_i = j \iff (\boldsymbol{\beta}_{i\phi_t}, \xi_i) = (\boldsymbol{\beta}_j^*, \xi_j^*)$. Let con-

sider a set of clusters $\{C_1, C_2, \dots, C_{K_I}\}$. To each cluster corresponds a distinct value for the parameters. For example, for cluster j , the cluster specific parameters are $(\boldsymbol{\beta}_j^*, \xi_j^*)$. In practice, the allocation variables describe the cluster assignments, that is $s_i = j \iff \mathbf{y}_i \in C_j$. Since we are clustering the \mathbf{y}_i , $i = 1, \dots, I$, each of which is a time series of length T , we obtain spatial clustering of time series.

The model is completed by specifying prior distributions for the model hyperparameters, σ_1^2 and τ^2 and ρ . The observational level variance σ_1^2 and the latent level scale parameter τ^2 are both assigned Inverse-Gamma priors. The choice of the prior on the parameter ρ is of critical importance, and we assign a beta prior to this parameter.

4 Posterior inference

Here, we focus on the identification of subregions of the metropolitan area of Milano which share similar temporal patterns. To do so, we report the Binder partition, which is the clustering of the areal units obtained by minimizing the posterior expected value of the Binder's loss function under equal misclassification costs ([2]). Briefly, the Binder loss function measures the distance for all possible pairs of subjects between the true probability of co-clustering and the estimated cluster allocation. Figure 2(a) shows the posterior co-clustering probabilities, i.e. the probability of every pair of areal units to be assigned to the same cluster, given the data; the numbering of the areal units is that given by the estimated Binder partition, which identifies 6 clusters. Figure 2(b) shows the Binder partition of the areal units overlapped with the region of interest. We can recover quite clearly the center of Milano (in purple), the first ring around the center (in red), and the most suburban areas (in yellow). We can also see that other areas with high concentration of activities (e.g., Rho), are clustered together with the first ring around the center (in red). The spatial association is very strong, as quantified through 95% credible interval of the marginal posterior distribution of ρ , equal to (0.897, 0.936).

5 Future work

In recent years, there is more interest in data about mobile networks usage. Mobile network providers, in fact, need to responsively detect poor network performances and Quality of Experiences issues correlated to the spatial and temporal distribution of the users. The goal is to take actions for their improvements, e.g. by changing network configuration to provide a better service in the places where users move or stay more. This is becoming more important for current 4G LTE mobile networks (and looking forward to next 5G networks), where the network configuration can be updated according to the discovered/predicted user mobility patterns. For these reasons, we hope to be able to apply our model and algorithms to 4G LTE data.

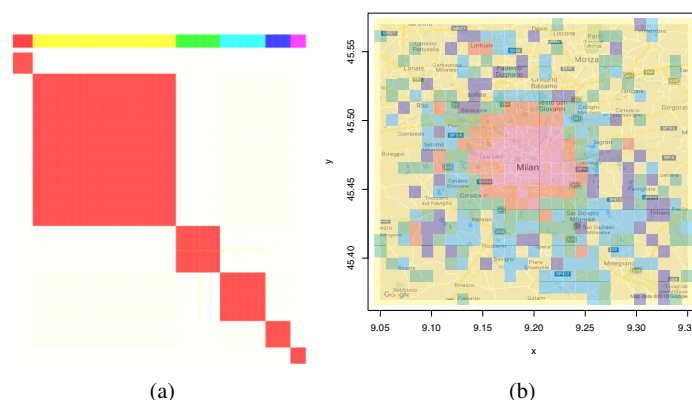


Fig. 2: Optimal clustering according to Binder's loss function – (a): posterior co-clustering probabilities. (b): estimated Binder partition of the areal units on the map of the area of interest in Milano.

Acknowledgements The authors would like to acknowledge prof. Simone Vantini, Politecnico di Milano, for supplying the data and for an interesting discussion.

References

1. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. *JRSS B*, 192–236 (1974)
2. Binder, D. A.: Bayesian cluster analysis. *Biometrika* **65**, 31–38 (1978)
3. De Iorio, M., Müller, P., Rosner, G. L., MacEachern, S. N.: An ANOVA model for dependent random measures. *JASA* **99**, 205–215 (2004)
4. Knorr-Held, L., Rue, H.: On block updating in Markov random field models for disease mapping. *Scandinav. J. of Statist.* **29**, 597–614 (2002)
5. Lee, D., Lawson, A.: Quantifying the spatial inequality and temporal trends in maternal smoking rates in Glasgow. *Ann. Appl. Statist.* **10**, 1427–1446 (2016)
6. Leroux, B. G., Lei, X., Breslow, N.: Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: *Statistical models in epidemiology, the environment, and clinical trials*, pp. 179–191. Springer, New York (2000)
7. Manfredini, F., Pucci, P., Secchi, P., Tagliolato, P., Vantini, S., Vitelli, V.: Treelet decomposition of mobile phone data for deriving city usage and mobility pattern in the Milan urban region. In: *Advances in complex data modeling and computational methods in statistics*, pp. 33–147. Springer, Cham, 2015.
8. McCausland, W. J., Miller, S., Pelletier, D.: Simulation smoothing for statespace models: A computational efficiency analysis. *Computational Statistics & Data Analysis*, **55**, 199–212 (2011)
9. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *JRSS,B* **71**, 319–392 (2009)
10. Secchi, P., Vantini, S., Vitelli, V.: Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. *Statistical Methods & Applications* **24**, 279–300 (2015)

A Mathematical Framework for Population of Networks: Comparing Public Transport of Different Cities.

Un approccio matematico all'analisi di una popolazione di networks: come confrontare il sistema di trasporto pubblico di diverse città.

Anna Calissano, Aasa Feragen, Simone Vantini

Abstract The analysis of relational data as network structures is very important in terms of modelling and applications. While the analysis of a single network is broadly studied, a rigorous and meaningful statistical modelling of population of networks is still little developed in the literature, in terms of Object Oriented Data Analysis. In this talk, we are going to describe our approach to populations of networks by applying and extending the Structure Space [10]. This space is a finite dimensional quotient space allowing networks to be the same up to permutations of nodes. We are going to introduce and compare different families of possible permutations, adapted to different application. A dataset of public transport networks of different cities around the world is analysed within this framework, allowing a comparison of multi-level networks with different nodes and topologies.

Abstract in Italiano *L'analisi di dati relazionali sta ricevendo sempre più attenzione da parte della comunità statistica. Mentre vi è molta letteratura che si concentra sull'analisi di un singolo network, i modelli che si occupano in maniera statisticamente rigorosa dello studio di popolazioni di networks sono ancora pochi. In questa relazione orale, descriveremo il nostro approccio allo studio di popolazioni di network basato sullo Structure Spaces [10], che permette l'analisi di popolazioni di networks a meno di permutazioni dei nodi. Estenderemo questo spazio considerando solo alcune sottopermutazioni di nodi con caratteristiche simili. Questo modello verrà raccontato analizzando un dataset contenente i network multilivello del trasporto pubblico di 25 diverse città nel mondo.*

Key words: Population of Networks, Network-value data, Quotient Space, Public Transportation System

Anna Calissano

MOX - Dept. of Mathematics, Politecnico di Milano e-mail: anna.calissano@polimi.it

Aasa Feragen

Dept. of Computer Science, University of Copenhagen e-mail: aasa@di.ku.dk

Simone Vantini

MOX - Dept. of Mathematics, Politecnico di Milano e-mail: simone.vantini@polimi.it

1 Motivation

In recent years, a soaring attention in the industrial and research community has been devoted to the analysis of tree-shape and network-shape data. According to the concept of first and second generation data analysis introduced by [19], the network analysis literature has mostly focused on the analysis of a network in a first generation setting (i.e. the analysis of a single network). The requirement of developing a second generation modelling for populations of trees and networks rises due to several applied problems, such as brain connectivity analysis [16, 3], blood vessel analysis [6], mobility networks analysis [18]. Populations of trees have been studied in the recent years under different perspectives. BHV space [1] as well as Tree-space [5] are examples of non-Euclidean tree-spaces in which tree-shape data can be embedded and analysed with several available tools. For what concerns network-shape data, there are currently two main approaches: Euclidean embedding and non-Euclidean embedding. Euclidean Graph embedding theory allows to embed networks in Euclidean space through different approaches such as kernel methods [15], convolutional neural networks [4], and feature selection algorithms [2]. If one is interested in clustering as well as predicting class of continuous variables from the embedded object, Graph Neural Embedding are powerful tools. However, they do not ensure that every point in the embedded space is actually a network. As well as they can assign non null probability to a set of points that are not networks. This causes problems in extending statistical tools to this embedding spaces. As a response to this problematic aspects, object oriented data analysis approach aims at building an embedding space that ensure that every point in the embedded space is an object. Some examples of this recent stream of literature are [8] and [10]. The latter approach introduces a quotient space, called Structure Space, where the orbit of the networks corresponds to all the equivalent networks up to permutations of nodes. We build our analysis on this latter approach, due to its flexibility in terms of applications. In this talk we are going to introduce the Structure Space, discuss its properties and the available statistical tools, considering the challenging aspect that Structure Space is a quotient space but not a smooth manifold. We are going to build on the Structure Space by introducing and allowing only certain types of node permutation. We are going to show the effect of changing the possible permutation on a real world dataset, collecting different cities' public transportation networks.

2 Data Set Description

In this talk, we are going to apply our techniques on a real world dataset collected and described in [13]. In the dataset, 25 different cities are selected, allowing a comparison of the mobility systems. The importance of complex system comparison is a key aspect for a better mobility system planning, a more conscious risk assessment, an interesting sociological analysis [17]. In this dataset, the public transportation system is collected for each city. The public transportation system accounts

for tram, rail, subway, ferry, bus, gondola, cablecar and funicular networks. Figure 1 shows the selected cities on the map. The dataset is complex and rich and allows for: between-city comparison of different transportation system, comparison of different transportation systems within each city, comparison of time evolving networks within and between each city. The data are collected during one week in December 2016.

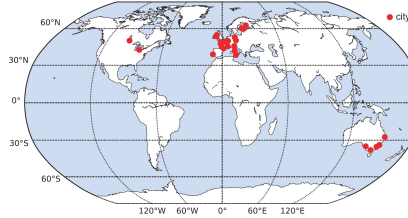


Fig. 1 Maps showing the 25 different cities analysed [13].

3 Talk Outline

Structure Space is a graph-space useful at describing populations of networks. In our perspective, networks are fully described by an *A-attributed r-structure* [10]. An *r-structure* is a pair $X = (P, R)$ consisting of a finite set P - e.g. the set of nodes in a network - and a subset $R \in P^r$ - e.g. the direct edges of the network when $r = 2$. The *r-structure* fully describes the topology of the network. An *A-attributed r-structure* is a triple $X_a = (P, R, a)$ consisting of a *r-structure* $X = (P, R)$ and an attribution $a : P^r \rightarrow A$, with $a(p)$ being non-null if, and only if, $p \in R$. The a map allows to describe attributes on both edges and nodes. As an example of an *A-attribute r-structure*, consider a city multi-layer public transportation network. P is the set nodes of the public transport system, i.e. the considered stops of the available transports system in the city. R is a subset of P^2 describing the edges between nodes (i.e. the presence of a trip between two stops). The attributes are the number of trips between stops in the existing transportation system in the city. A is \mathbb{R}^s where $s \leq 7$, i.e. the number trips of seven different measured transportation systems. Note that loops in this case are not allowed. In many real world applications, populations of networks are measuring the same phenomenon for different specific contexts. Due to different nodes' labels or order, topological similarities between network could not straightforwardly be investigated. For example, public transportation stops are labelled differently in different cities so a certain type of node permutation should be allowed (e.g. we would like to be able to identify the center stations among cities, the nightlife areas, etc.).

Allowing permutation corresponds to defining the following isomorphism. Two A -attributed r -structures $X_1 = (P_1, R_1, a_1)$ and $X_2 = (P_2, R_2, a_2)$ are isomorphic, written as $X_1 \sim X_2$, if there is a mapping $f : P_1 \rightarrow P_2$ satisfying:

1. $p = (p_1, \dots, p_r) \in R_1 \Leftrightarrow f(p) = (f(p_1), \dots, f(p_r)) \in R_2$
2. $a_1(p) = a_2(f(p))$ for all $p \in R$.

Networks can be studied in the quotient space obtained by identifying each network with its entire equivalence class, consisting of all possible networks obtained by permuting the original nodes. Note that this quotient space is not a manifold due to the fact that the permutation group is not acting freely [14], so all the statistical tools extended on manifold are not available here (see [9, 11, 12] for developed techniques on this quotient space).

Choosing the allowed class of permutations f is a crucial point in the analysis of networks. In this talk we are going to describe in details how this quotient space can be built, by allowing different type of permutations. We are going to propose a novel permutation approach allowing only certain permutations, according to node characteristics. Changing the quotient space by changing the possible sub-group of permutations implies very different results in terms of analysis. In addition, by working only on sub-permutation, computational cost of network matching can be decreased and scalability consequently improved. The population of public transportation systems described in Section 1 are studied allowing different type of permutation, thus different embedding space. In the different spaces, we compute basic statistics such as Fréchet mean [7] and cluster analysis. Different results will be shown, revealing the importance of the space the networks are embedded in.

Acknowledgements This project is partially funded by the *SAFARI NJEMA : From informal mobility to mobility policies through big data analysis*. Polisocial Award 2018, Politecnico di Milano (<http://www.polisocial.polimi.it/it/home/>).

References

- [1] Louis J Billera, Susan P Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.
- [2] Horst Bunke and Kaspar Riesen. Improving vector space embedding of graphs through feature selection algorithms. *Pattern Recognition*, 44(9):1928–1940, 2011.
- [3] Daniele Durante, David B Dunson, and Joshua T Vogelstein. Nonparametric bayes modeling of populations of networks. *Journal of the American Statistical Association*, 112(520):1516–1530, 2017.
- [4] David K. Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Ad-*

- vances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2224–2232, 2015.
- [5] Aasa Feragen, Francois Lauze, Pechin Lo, Marleen de Bruijne, and Mads Nielsen. Geometries on spaces of treelike shapes. In *Asian Conference on Computer Vision*, pages 160–173. Springer, 2010.
- [6] Aasa Feragen, Megan Owen, Jens Petersen, Mathilde MW Wille, Laura H Thomsen, Asger Dirksen, and Marleen de Bruijne. Tree-space statistics and approximations for large-scale analysis of anatomical trees. In *International Conference on Information Processing in Medical Imaging*, pages 74–85. Springer, 2013.
- [7] Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l’institut Henri Poincaré*, 10(4):215–310, 1948.
- [8] Cedric E. Ginestet, Jun Li, Prakash Balachandran, Steven Rosenberg, and Eric D. Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.*, 11(2):725–750, 06 2017.
- [9] Brijnesh Jain and Klaus Obermayer. On the sample mean of graphs. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 993–1000. IEEE, 2008.
- [10] Brijnesh J Jain and Klaus Obermayer. Structure spaces. *Journal of Machine Learning Research*, 10(Nov):2667–2714, 2009.
- [11] Brijnesh J Jain and Klaus Obermayer. Large sample statistics in the domain of graphs. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 690–697. Springer, 2010.
- [12] Brijnesh J Jain and Klaus Obermayer. Learning in riemannian orbifolds. *arXiv preprint arXiv:1204.4294*, 2012.
- [13] Rainer Kujala, Christoffer Weckström, Richard K Darst, Miloš N Mladenović, and Jari Saramäki. A collection of public transport network data sets for 25 cities. *Scientific data*, 5:180089, 2018.
- [14] John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–31. Springer, 2013.
- [15] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561, 2011.
- [16] Sean L Simpson, Robert G Lyday, Satoru Hayasaka, Anthony P Marsh, and Paul J Laurienti. A permutation testing framework to compare groups of brain networks. *Frontiers in computational neuroscience*, 7:171, 2013.
- [17] J Michael Thomson. *Great cities and their traffic*. Penguin Books, 1978.
- [18] Christian von Ferber, Taras Holovatch, Yu Holovatch, and V Palchykov. Public transport networks: empirical analysis and modeling. *The European Physical Journal B*, 68(2):261–275, 2009.
- [19] Haonan Wang, JS Marron, et al. Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35(5):1849–1873, 2007.

How Important Discrimination is for the Job Satisfaction of Immigrants in Italy: A Counterfactual Approach

Quanto influisce la discriminazione sulla soddisfazione lavorativa degli immigrati in Italia: un approccio controfattuale

Maria Gabriella Campolo, Antonino Di Pino and Michele Limosani

Abstract In this study we aim to evaluate how the status of citizenship, for the immigrants in Italy, influences their job satisfaction and perception of well-being. To this end, analysing the Eusilc-Italy data, we try to disentangle the effect of discrimination due to the citizenship status from the effect of the individual skills and socio-demographic characteristics on the individual job satisfaction and life satisfaction. To this end we follow a counterfactual approach by applying the Blinder-Oaxaca decomposition. As a result, we found that individual skills determine differences in earning work income between immigrants and natives, but do not significantly explain the different perception of well-being and job satisfaction, probably due to a discrimination effect.

Abstract L'obiettivo di questo studio è valutare quanto trovarsi nella condizione di immigrato in Italia influenzi la soddisfazione per il proprio lavoro e la percezione del proprio benessere. A tale scopo, analizzando i dati di Eusilc-Italia, proponiamo un metodo basato sulla decomposizione di Blinder-Oaxaca per identificare l'effetto sulla percezione del benessere della discriminazione dovuta allo status di immigrato rispetto all'effetto prodotto dalle capacità individuali e da altri fattori di natura socio-economica e demografica. Ne deriva che le abilità individuali determinano differenze nel reddito da lavoro tra immigrati e nativi, ma non spiegano in modo significativo il diverso livello di soddisfazione nel contesto lavorativo, da attribuirsi verosimilmente ad un effetto di discriminazione.

Key words: Job Satisfaction, Life Satisfaction, Blinder-Oaxaca Decomposition

¹

Maria Gabriella Campolo, Università degli Studi di Messina; email: mgcampolo@unime.it

Antonino Di Pino, Università degli Studi di Messina; email: dipino@unime.it

Michele Limosani, Università degli Studi di Messina; email: limosani@unime.it

Introduction

In the last years, analysts paid particular attention to the factors that influence the immigrants assimilation in the socio-economic context of developed countries. In particular, the comparison between the socioeconomic condition of immigrants and that of natives has been extensively studied (see, among others, Van Tubergen et al. 2014). However, a relevant theme of research about the immigrants condition, given by the analysis of the perception of their subjective well-being and position in society, can still be further developed. A relevant contribution has been recently provided by Arpino and De Valk (2018) who emphasize the role of the social embeddedness as an explanatory factor for life satisfaction of immigrants.

With regard to studies on the relationship between economic resources and well-being, analysts (e.g. Easterlin, 2001; De Jong et al., 2002) found that, in societies with advanced development, an inverse correlation between income and satisfaction for life occurs (so-called Easterlin's paradox), being included in satisfaction for life also that for the working condition or job satisfaction.

Therefore, if these studies show that, in general, "money does not buy happiness", a counterintuitive implication arises for the immigrant category: that is, migrants are wrong in believing that economic migration is a way to improve one's own well-being, at least to the extent that well-being means (or includes) happiness. However, other studies have shown (e.g. Bartram, 2011) that the correlation between the level of satisfaction with life and income is negative for natives and positive for immigrants. This implies that migrants could represent an exception to Easterlin's paradox, deriving their wellbeing (happiness) condition from increased incomes to a greater extent than native people.

Focusing on the theme of the native-immigrant comparison, Arpino and De Valk (2018) empirically measure the satisfaction of life of immigrants and native throughout Europe. Using the data of the first six rounds (2002-2012) the European Social Survey, the authors find that the life satisfaction among immigrants is lower than that of the natives, although the differences diminish over the generations. For second generation immigrants, for example, a relevant part of the life satisfaction gap is explained by the lower level of social membership than natives.

With regard to the present analysis, focused specifically on the first generation immigrants (regular for legal status) behavior, we would investigate the extent to which different levels of life satisfaction and work satisfaction between immigrants and natives are depending on the different skills of the subjects, or, otherwise, are depending on the membership of the subject in one of the two sub-groups (native or immigrants).

To this end, we perform an empirical analysis using the data provided by Istat as part of the Eusilc Survey promoted in 2013 for Italy. The data of the Eusilc surveys are available on request to Istat since 2004, however a module regarding the satisfaction level information set ("Wellness Module") has been included only for the 2013 survey (this prevents us to provide an over-time differential analysis).

In order to evaluate to what extent the level of job satisfaction is due to the influence of standard variables such as income, job position, education, health, etc.,

How Important is Discrimination on the Job Satisfaction of Immigrants in Italy

and to what extent, instead, depends on the different status of native or immigrant, an original application of the model of decomposition proposed by Blinder (1973) and Oaxaca (1973) (B.O. decomposition) is here suggested. Specifically, we adopted a more recent version of the B.O. model, improved for more complex analysis implying sample selectivity due to censored observations (e.g. Jann, 2008).

As a result of this analysis, we found that differences in demographic and socio-economic variables that distinguish the two groups (age, education, type of work, etc.) of immigrants and Italian natives, respectively, do not explain convincingly the lower level of job satisfaction and well-being of immigrants. This result leads us to believe that a latent discrimination effect explains the difference in perceived wellbeing between the two groups.

Data and Methods

The B.O. decomposition allows us to evaluate whether the differences in job and life satisfaction between immigrants and Italian natives can be assessed by differences in the observable characteristics of the subjects (so called “endowments” effect) or by the structural change in coefficients between the two groups, respectively, of immigrants and natives (so-called “coefficients” effect). The latter effect can be interpreted as an effect of “discrimination” due to the belonging of the subject to one or the two groups. In this case, one of the two groups would be “discriminated” and, on average, the analyzed outcome (i.e. the job satisfaction) in a group is lower than that in the other group because of the discrimination. Another effect (so-called “interaction” effect) is given by the interaction between the first two effects, endowments and coefficients.

If, from the results of the decomposition, the “coefficients” component prevails, this implies that the effect of covariates measuring endowment is less relevant. In this context the effectiveness of integration policies diminishes, being these policies usually based on intervention programs that modify the covariates related to the individual wellbeing (e.g. income integrations and subsidies, interventions on education, health services, kindergartens, etc.). At the opposite, if the endowments component prevails, this implies that only by influencing the covariates measuring the individual skills through ad hoc policies, we can reduce over time both the differential in wellbeing and in perceived satisfaction between subjects belonging, respectively, to the less-endowed and to more-endowed groups.

Following the methodology of Oaxaca (1973) and Blinder (1973), we introduce a Three-fold decomposition of the difference in mean outcomes given by the surveyed scores of both life satisfaction and job satisfaction of, respectively, Italian natives (N) and immigrant (I). This outcome decomposition, denoted R , is specified as follows:

$$R = M(Y_N) - M(Y_I) = (\bar{X}_N - \bar{X}_I)\hat{\beta}_I + \bar{X}_I(\hat{\beta}_N - \hat{\beta}_I) + (\bar{X}_N - \bar{X}_I)(\hat{\beta}_N - \hat{\beta}_I) \quad (1)$$

Where $M(Y)$, denotes the mean of the outcome variable for two different groups (in our analysis natives N and immigrants I). The first term of Eq.1, $E = (\bar{X}_N - \bar{X}_I)\hat{\beta}_I$, represents the Endowments effect and explains the differences due to covariates characteristics (such as education, age, sex, the health condition of the subject, the sector of economic activity). The second term, $C = \bar{X}_I (\hat{\beta}_N - \hat{\beta}_I)$ represents the Coefficients effect (C), and explains structural changes in the coefficients of the covariates between the two groups. Finally, the third term, Interaction effect (I) given by $I = (\bar{X}_N - \bar{X}_I)(\hat{\beta}_N - \hat{\beta}_I)$, allow us to take into account the fact that differences in endowments and coefficients between natives and immigrant exist simultaneously. This term shows how the predicted outcome of immigrant would have changed if he/she had the same observable and unobservable characteristics as the native.

When both groups obtain equal returns for their characteristics, the outcome differentials will be explained by the difference in endowments alone (the second and the third part of the decomposition will equal zero). The above decomposition is formulated based on the prevailing outcome (life or job satisfaction level) of natives, while the differences in endowments and coefficients between immigrants and natives are weighted by the satisfaction coefficients of immigrants.

A methodological problem arises in our analysis due to the fact that the outcome variables, job satisfaction and life satisfaction (to a lesser extent, but still relevant, also the yearly labour income presents censored data). For this reason, we have adopted a variant of the B.O. decomposition model that takes into account the censoring in the outcome variable (e.g., Jann, 2008). For the direct application of models of decomposition with sample selectivity correction due to censoring in the outcome variable, the STATA14 package proposed by Jann (2008) has been adopted.

As for the sample characteristics, we draw a sample from the 2013 EU-SILC module for Italy, conducted by the National Institute of Statistics (ISTAT) that include about 29,000 households (about 70,000 individuals). For the present analysis, we select a sample of Italian natives and regular immigrants composed by 11467 subjects, aged 16-70 and active on the labor market. In the selection of the sample we tried to represent faithfully (as far as possible), with respect to official statistics, the distribution by gender and by region of residence, as well as the distribution of the respondents by nationality. In this way, the sample consists of 10262 natives and 1205 foreigners.

Decomposition Results and Discussion

The results of the B.O. decomposition suffer for the influence of a great number of non-responses and for the consequent sampling selectivity. The most impact of censoring was found for the job-satisfaction score answers. From the decomposition of labor income, whose differential is unfavorable to immigrants, we found a prevalence of a significant "endowments" effect, that is the component due to

How Important is Discrimination on the Job Satisfaction of Immigrants in Italy

differences in socio-economics covariates that affect income (see Table 1). In particular, the difference of 1309,6 between natives and immigrants due to endowments effect indicates that differences in age, education level, gender, health status, geographical area, hours worked and type of job account for about 87% the wage gap.

At the opposite a significant gap in job satisfaction, unfavourable to immigrants is prevalently due to the “coefficients” effect (equal to 0.53). This result may be attributable to a structural change in the coefficients between the two groups caused by latent factors, probably related to the sphere of social inclusion and to the perception of own status. Finally, considering the level of life satisfaction, the difference in favour of natives is equal to 0.61 (7.76 for natives and 7.14 for immigrants). In particular, the gap favourable to natives is prevalently due to the structural change in coefficients (equal to 0.43) and to the interaction effect between endowments and coefficients effect (equal to 0.35).

Table 1: Estimation results of Oaxaca-Blinder three-fold decomposition of yearly labour income, job satisfaction, life satisfaction in Italy

Dependent variable:	Yearly labour income			Job Satisfaction			Life satisfaction		
	Coef.	Std. Err.	p.value	Coef.	Std. Err.	p.value	Coef.	Std. Err.	p.value
<u>OVERALL</u>									
Natives - group 1	16141.8	84.6	***	7.64	0.16	***	7.76	0.15	***
Immigrants- group 2	14673.7	255.6	***	7.02	0.14	***	7.14	0.45	***
Difference	1468.1	269.3	***	0.62	0.22	**	0.61	0.47	
<i>Decomposed by effect of:</i>									
endowments	1309.6	266.1	***	0.00	0.11		-0.17	0.35	
coefficients	94.6	54.2		0.53	0.24	*	0.43	0.48	
interaction	63.9	45.7		0.09	0.15		0.35	0.37	
<u>ENDOWMENTS</u>									
Yearly Labour income				-0.41	0.29		6.44	23.53	
Area ^a	-601.1	99.9	***	-0.11	0.11		3.75	13.55	
Age	568.7	57.1	***	0.10	0.11		-1.57	4.22	
Weekly working hours	758.4	131.2	***	0.26	0.17		-3.89	14.01	
Education ^a	60.7	67.3		0.01	0.03		-0.33	1.16	
Permanent employment ^a	282.2	128.7	*	0.13	0.09		-0.24	0.71	
Woman ^a	238.8	83.6	**	0.04	0.05		-0.52	1.40	
Health status	1.9	3.7					-0.04	0.32	
Constant				-0.02	0.02				
<u>COEFFICIENTS</u>									
Yearly Labour income				2.55	3.55		-133.15	249.64	
Area ^a	-102.3	34.2	*	-0.27	0.50		24.24	45.33	
Age	-75.2	156.2	***	-0.18	1.38		25.21	38.71	
Weekly working hours	-222.7	35.0	***	-0.37	0.68		26.23	48.63	
Education ^a	2443.4	139.2	***	-0.42	1.63		48.01	85.49	
Permanent employment ^a	-52.5	19.8	**	0.35	0.27		-2.67	4.19	
Woman ^a	-544.3	45.9	***	0.15	0.46		-5.92	9.43	
Health status	52.8	79.1		0.06	0.49		2.54	6.42	
Constant	-1404.6	251.3	***	-1.35	1.63		-895.4	1788.7	
<u>INTERACTION</u>									
Yearly Labour income				0.28	0.40		-13.03	24.59	
Area ^a	31.8	11.8	**	0.08	0.15		-7.52	14.15	
Age	-8.8	18.3		-0.02	0.16		2.85	4.39	
Weekly working hours	-68.1	15.5	***	-0.12	0.22		7.85	14.63	
Education ^a	28.3	31.4		0.00	0.01		0.64	1.35	

Table 1 - continued

Permanent employment ^a	10.5	6.1		-0.12	0.11	0.46	0.79
Woman ^a	67.5	24.1	**	-0.02	0.06	0.91	1.48
Health status	2.6	4.1		0.00	0.03	0.13	0.33
No. of natives	6019			2206		4264	
No. of immigrants	556			174		374	

Note: p. value: *** <0.001; ** < 0.01; * < 0.05; ^a Area (Dummy:1 = North_Centre), Education (years of schooling), Permanent employment (Dummy: 1=permanent), Woman (Dummy:1=yes)

In sum, different conclusions derived from the decomposition of, respectively, labour income and the score measuring the job satisfaction and the life satisfaction. In the first case, the differences between natives and immigrants depend on socio-economic factors that influence the production of income. At the opposite, the differences in job satisfaction and life satisfaction seem not to depend so prevalent on the socio-economic covariates measuring the individual skills, but rather on a structural change of coefficients interpretable as an effect of discrimination.

Compared to the results of the authors mentioned above, the findings of our analysis seem to support the conclusion of the Arpino and De-Walk (2018) study, that is differences between natives and immigrants in the perception of their wellbeing cannot be easily explained by the influence of economic and demographic factors and of their own individual skills. In this context we can assume that, if income, education, health and other endowment factors do not give happiness, it might be worth the Easterlin paradox: that is, the individuals attributes value rather than to the absolute level of their economic well-being, to their own status within the community and the social recognition of it.

References

1. Arpino, B., de Valk, H. : Comparing life satisfaction of immigrants and natives across Europe: The role of social contacts. Soc Indic Res. (2018) doi: 10.1007/s11205-017-1629-x
2. Bartram, D.: Economic migration and happiness: Comparing immigrants' and natives' happiness gains from income. Soc Indic Res. (2001) doi: <http://dx.doi.org/10.2139/ssrn.1590265>
3. Blinder, A. S.: Wage discrimination: reduced form and structural estimates. J. Hum. Resour. (1973) doi: 10.2307/144855.
4. De Jong, G. F., Chamratrithirong, A., Tran, Q.G.: For better, for worse: Life satisfaction consequences of migration. Int. Migr. Rev. (2002) doi: <https://doi.org/10.1111/j.1747-7379.2002.tb00106.x>
5. Easterlin, R. A.: Income and happiness: Towards a unified theory. Econ. J. (2001) doi: <https://doi.org/10.1111/1468-0297.00646>
6. Jann, B.: The Blinder–Oaxaca decomposition for linear regression models. Stata J. (2008) doi: <https://doi.org/10.1177/1536867X0800800401>
7. Oaxaca, R.: Male-female wage differentials in urban labor markets. Int. Econ. Rev. (1973). 14(3): 693-709 doi: 10.2307/2525981.
8. Van Tubergen, F.: Size and socio-economic resources of core discussion networks in the Netherlands: Differences by national-origin group and immigrant generation. Ethn. Racial Stud. (2014) doi: <https://doi.org/10.1080/01419870.2012.734390>

Unfolding the SEcrets of LongEvity: Current Trends and future prospects (SELECT)

A path through morbidity, disability and mortality in Italy and Europe

Stefano Campostrini, Daniele Durante, Fabrizio Faggiano and Stefano Mazzuco

Abstract The rapid social, economic, and technological transformation characterizing our society in the recent years is producing several effects on many complex and dynamic processes of human health. We focus, in particular, on the recent upward trend in longevity, and on its relations with current and future morbidity and disability patterns. The joint analysis of these processes, which play a key role in many public health systems, requires novel qualitative and quantitative paradigms. The focus of SELECT is to unfold the secrets of longevity and learn dynamic interrelations between morbidity, disability and mortality, while identifying future trends in order to devise effective interventions under a collective effort by demographers, epidemiologists, social and data scientists.

Abstract *Le rapide trasformazioni che hanno caratterizzato nei recenti anni la nostra società, stanno producendo diversi effetti sulle dinamiche complesse della salute. Ci concentriamo, in particolare, sull'aumento della longevità e sul suo rapporto con i trend futuri di morbidità e disabilità. L'analisi congiunta di questi processi, che giocano un ruolo determinante nei sistemi di salute pubblica, richiede infatti nuovi approcci qualitativi e quantitativi. Nel progetto SELECT, si vogliono studiare i segreti della longevità, scoprendo le interrelazioni dinamiche tra morbidità, disabilità e mortalità, per elaborare interventi efficaci. Tali obiettivi richiedono lo sforzo collettivo di demografi, epidemiologi, statistici sociali e data scientists.*

Key words: data science, demography, epidemiology, public health, statistics

Stefano Campostrini
University Ca' Foscari Venice, e-mail: stefano.campostrini@unive.it

Daniele Durante
Bocconi University, e-mail: daniele.durante@unibocconi.it

Fabrizio Faggiano
Università del Piemonte Orientale, e-mail: fabrizio.faggiano@uniupo.it

Stefano Mazzuco
University of Padova, e-mail: stefano.mazzuco@unipd.it

1 Removing barriers for an improved understanding of longevity, disability and morbidity processes

Italy is acknowledged as one of the countries with the highest longevity in the world. Evidence of this is provided not only by mortality statistics (OECD, 2017), but also by several studies focusing on high prevalence of super-centenarians in some regions (e.g. Poulain et al., 2004). This phenomenon is expected to have a considerable impact on public health and sustainability, especially if increasing longevity is not accompanied by a similar decrease in morbidity and disability, as happened over the recent years (ISTAT, 2017). Despite the relevance of the topic, there are few studies aimed at understanding the complex patterns underlying these processes, their determinants and possible future patterns. Indeed, there are three main barriers, discussed below, to the progress in such a field.

1. **DATA:** Although there is a wide availability of separate sources of information on morbidity, disability and mortality, few attempts have been made to link these datasets in order to work on a wider picture of the whole process. In fact, such phenomena are highly related and a joint multidisciplinary analysis could yield to substantial advances in this field, well beyond the findings provided by separate analyses. Besides linkage problems, there are also measurement issues, leading to debated literature and findings. For instance, Guillot and Canudas-Romo (2016) point out that some countries characterized by a fast mortality decline show a period mortality lower than others, but lose rankings when a cohort perspective is taken. Thus, we are faced with a dilemma on whether to use timely but possibly misleading period measures, or more appropriate but untimely cohort ones.
2. **STATISTICAL MODELS:** There have been substantial advances in statistics and data science over the recent years (e.g. Donoho, 2017). However, most of the motivations underlying new learning methods typically come from genetics, neuroscience, engineering, and precision medicine, rather than morbidity, disability and mortality modeling. Closing this gap would lead to major improvements in understanding longevity. Indeed, new models that can explain how future mortality curves change across states or vary as a function of causes-of-deaths, incidence of diseases and prevalence of risk factors, may suggest key guidelines for healthy longevity. Also disability, and its relations with morbidity and mortality, is gaining substantial attention, and could benefit by improved modeling, thus answering conflicting literature on this topic (e.g. Christensen et al., 2008; Angel et al., 2015). A major issue will be that of taking into consideration evolutionary and spatial aspects (Assaf et al., 2016; Assaf and Campostrini, 2014).
3. **EVALUATION:** Given the current financial constraints in public health, policy-makers have to select interventions that are effective, efficient and as timely as possible. But no instruments are currently available to support this complex and crucial task (Faggiano et al., 2014). However, the complexity of the interactions between risk factors, incidence of diseases and healthy longevity need to be considered, so that all direct and indirect effects of the strategy can be properly eval-

SELECT

uated. This requires the development of comprehensive modeling tools able to simulate the impact of a “bouquet” of different policies, selected on the base of their effectiveness, transferability and sustainability, on the prevalence of risk factors, on the incidence of diseases and on the final effect on mortality, considering also sub-populations composed by social categories. To provide successful outcomes, this evaluation requires deep interaction and constant feedbacks, currently limited, between experts from different fields.

Removing the above mentioned barriers requires substantial advances in data reconstruction, together with the development of innovative statistical models for complex morbidity and longevity patterns, combined with a constant exploitation and reference to current public health or epidemiological theories. In this paper, we describe the SELECT research project that has been recently granted by MIUR (“Ministero Italiano dell’Istruzione, dell’Università e della Ricerca”) among PRIN (Research projects of national interest) projects. The aim is to take a relevant step to tackle the above mentioned issues under different research directions motivated by a multidisciplinary team, and by the availability of several sources of information, including, among the others, the Human Mortality Database, the Human Cause-of-Death Database, the morbidity and risk factor surveillance system PASSI (Baldissera et al., 2011), the Global Burden of Disease, the Cochrane Library, the Health Evidence database, and the NICE guidance for the summaries of effects of public health interventions, providing information on disability across countries, and others.

Consistent with the above considerations, SELECT will be articulated into three highly interconnected streams of research.

1. **DATA RECONSTRUCTION AND DATA LINKAGE:** Motivated by the absence of systematic data reconstruction methods for cohort mortality and by the lack of joint databases on morbidity, disability and mortality, a first step of this project will be on developing new strategies for data reconstruction and data linkage that allow the definition of new models able to provide a broader picture for the complex components of longevity.
2. **MODELING AND PREDICTING COMPLEX PATTERNS OF MORTALITY, MORBIDITY RISK FACTORS AND DISABILITY:** The research on data reconstruction is expected to provide a wide variety of rich datasets that combine information from a large range of databases and motivate the development of novel statistical models in different directions. Indeed, morbidity, disability and mortality are complex and dynamic processes characterized by several underlying patterns of interrelationships, thus requiring substantial advances in the modeling techniques, well beyond the available ones. This is particularly true in dynamic modeling and forecasting of mortality curves across countries and causes-of-death. These advances will arise at the intersection between functional data analysis, hierarchical modeling, filtering and dynamic inference for multivariate and mixed-scale processes, and structural equation models along with more flexible Bayesian nonparametric strategies and density regression methods.

3. SIMULATION AND EVALUATION OF INTERVENTION STRATEGIES: The rich databases reconstructed and the novel statistical models are expected to shed light into several complex patterns underlying longevity processes, their future trends and their relations with risk factors. This will open new avenues toward improving intervention strategies for healthy aging. Clearly, novel interventions require careful evaluation and a thorough assessment in the light of current strategies and prevention policies. We address this aim by proposing novel simulation methods which will be developed in constant interaction with the intervention policies provided by our quantitative findings and with those suggested by a careful review of the available literature on healthy aging.

A peculiarity of the research group is its multidisciplinary composition, which fosters integration and collaboration, especially for these cross-cutting themes. This will be further stimulated by organizing regular meetings of the steering committee, workshops, conferences and by developing online infrastructures.

References

1. Angel RJ et al (2015) Longer lives, sicker lives? *J Gerontol: Series B*, 70:639-649
2. Assaf S, Campostrini S (2014) Application of the varying coefficient model to the behaviour risk factor surveillance data in Italy. *BMC Public Health*: 489
3. Assaf S, Campostrini S et al (2016) Analysing behavioural risk factor surveillance data by using spatially and temporally varying coefficient models. *J R Stat Soc Ser A*, 179: 153–175
4. Baldissera S, Campostrini S et al (2011) Features and initial assessment of the Italian Behavioral Risk Factor Surveillance System (PASSI), 2007-2008. *Prev Chronic Dis*, 8:1–8
5. Christensen K et al (2008). Exceptional longevity does not result in excessive levels of disability. *PNAS*, 105:13274–13279
6. Donoho D (2017) 50 Years of Data Science. *J Comput Grap Statist*, 26:745–766
7. Faggiano F, Allara E et al (2014) Europe Needs a Central, Transparent, and Evidence-Based Approval Process for Behavioural Prevention Interventions. *PLoS Med*: e1001740
8. Guillot M & Canudas-Romo V (2016) Revisiting life expectancy rankings in countries that have experienced fast mortality decline. In Schoen R *Dynamic Demographic Analysis*, pp.51–67.
9. ISTAT (2017) BES 2017: il benessere equo e sostenibile in Italia. Istat, Roma
10. OECD (2017) Health at a Glance 2017: OECD Indicators. OECD Publishing, Paris.
11. Poulain M et al (2004) Identification of a geographic area characterized by extreme longevity. *Exp Gerontol*, 39:1423–1429

Galaxy color distribution estimation via dependent nonparametric mixtures

Stima della distribuzione del colore delle galassie via misture nonparametriche dipendenti

Antonio Canale, Riccardo Corradin and Bernardo Nipoti

Abstract We analyse a dataset drawn from the Sloan Digital Sky Survey first data release and consisting of the measurement of the $u - r$ color (i.e. difference between ultraviolet and red magnitude) of 24 312 galaxies. Galaxies are stratified by means of available covariate information on their luminosity and environment. An accurate estimate of the color distribution within each stratum provides interesting insight on how the population of galaxies changes across different luminosity and environment conditions. We propose and implement a Bayesian nonparametric mixture model for partially exchangeable data, obtained by normalising dependent superposed gamma processes. The structure of the proposed model allows for stratum-wise posterior inference while favouring borrowing of information across strata sharing same luminosity or same environment.

Abstract Analizziamo un dataset tratto dalla prima distribuzione di dati Sloan Digital Sky Survey e composto da misure di colore $u - r$ (cioè differenza tra magnitudine ultravioletta e rossa) di 24 312 galassie. Le galassie sono stratificate in base alle covariate luminosità e ambiente. Una stima accurata della distribuzione del colore all'interno di ogni strato permette di comprendere meglio come la popolazione di galassie cambi al variare di luminosità e ambiente. Proponiamo ed implementiamo un modello mistura bayesiano nonparametrico per dati parzialmente scambiabili, ottenuto normalizzando misture dipendenti di processi gamma. La struttura del modello proposto permette di fare inferenza specifica per ciascuno strato e, allo stesso tempo, favorisce scambio di informazione tra strati caratterizzati da stessa luminosità o stesso ambiente.

Antonio Canale

Dipartimento di Scienze Statistiche, Università degli studi di Padova, e-mail: canale@stat.unipd.it

Riccardo Corradin

DEMS, Università degli studi Milano-Bicocca, e-mail: riccardo.corradin@unimib.it

Bernardo Nipoti

School of Computer Science and Statistics, Trinity College Dublin, e-mail: nipotib@tcd.ie

Key words: Bayesian nonparametrics, dependent Dirichlet processes, mixture models, partially exchangeable data, galaxy surveys.

1 Astronomical application

Galaxies are known to be roughly composed by two populations, termed late-type and early-type galaxies. While this distinction originally refers to morphological properties of galaxies, the two populations can be approximately identified by their properties in terms of galaxy colors, such as, for instance, the $u - r$ color (i.e. difference between ultraviolet and red magnitude), with a bimodal distribution for the $u - r$ color reflecting the existence of two populations. Note that the definition of $u - r$ is such that larger values of $u - r$ correspond to redder colors. Considering that quantifying the $u - r$ color of galaxies is much simpler than assessing their morphological properties, considerable attention has been dedicated in the astronomical literature to the estimation of the color distribution and to the effect of available covariates on such distribution (see [1] and references therein). For instance, the $u - r$ color distribution is known to depend on both luminosity and galaxy environment. The luminosity is a measure of the light emitted by the galaxy in a given spectral band; the environment is quantified by measuring, for each galaxy, the density of neighbouring galaxies. In this paper we consider a dataset referring to $n = 24\,312$ galaxies, drawn from the Sloan Digital Sky Survey¹ (SDSS) first data release. Following [1], we consider a stratification of the SDSS dataset into 25 groups, based on five different levels of r -band luminosity and environment. Our goal is to accurately estimate the $u - r$ color distribution within each stratum so to provide interesting insight on the properties of the two main populations of galaxies and on how such properties change across strata.

Before describing our modelling framework, some notation is introduced. Let X_i , with $i = 1, \dots, n$, be the observed measurement of the $u - r$ color of a galaxy and let $Z_{1,i}$ and $Z_{2,i}$ be the related measures of its luminosity and environment. Both covariates are assumed discrete and take values in $\mathcal{L} = \{1, \dots, L\}$ and $\mathcal{G} = \{1, \dots, G\}$, respectively. Consistently with [1], here $L = G = 5$. This structure induces a stratification of the observations, which can be gathered into 25 distinct groups $\mathbf{X}^{(l,g)}$, with $l \in \mathcal{L}$ and $g \in \mathcal{G}$, where

$$\begin{aligned}\mathbf{X}^{(l,g)} &= \{X_i : Z_{1,i} = l, Z_{2,i} = g, 1 \leq i \leq n\} \\ &= \{X_1^{(l,g)}, \dots, X_{n_{l,g}}^{(l,g)}\},\end{aligned}$$

and $\sum_{l \in \mathcal{L}} \sum_{g \in \mathcal{G}} n_{l,g} = n$.

¹ <https://www.sdss.org>, for project details and relevant references

2 Modelling framework

While homogeneity can be reasonably assumed within each group, distinct groups are characterized by heterogeneous luminosity and environment conditions. At the same time, any sound model for the SDSS data should account for the fact that, for any fixed l (or g), the groups of observations $\{\mathbf{X}^{(l,1)}, \dots, \mathbf{X}^{(l,G)}\}$ (or $\{\mathbf{X}^{(1,g)}, \dots, \mathbf{X}^{(L,g)}\}$) are characterized by common luminosity (or environment). This suggests that a model for partially exchangeable data could be adopted so to allow for stratum-wise posterior inference while favouring borrowing of information across strata sharing same luminosity or same environment.

2.1 Two-level GM-Dependent Dirichlet processes

The last two decades have witnessed a growing interest in models for partially exchangeable data. The introduction of the celebrated dependent Dirichlet process by MacEachern in [8, 9] paved the way to the definition of a plethora of nonparametric models inducing borrowing of information across groups of observations, see [3] for a review. Here we consider a class of mixture models with dependent random mixing probability measures, obtained by normalizing dependent superposed gamma completely random measures (CRM), in the spirit of [5]. Specifically, we consider the class of Griffiths-Milne dependent Dirichlet processes (GM-DDP), defined and studied in [6, 7] and we adapt it so to accommodate the specific nature of the SDSS data. Considering the two possible sources of common randomness across groups, namely common density or common environment, we refer to the resulting vector of random probability measures as to *two-level* GM-DDP. For an allied approach see [4], for a detailed introduction to CRMs see [2].

In order to define a vector of two-level GM-DDP, we consider independent sets of independent gamma CRMs, namely $\{\mu_{l,0} : l \in \mathcal{L}\}$, $\{\mu_{0,g} : g \in \mathcal{G}\}$ and $\{\mu_{l,g} : l \in \mathcal{L}, g \in \mathcal{G}\}$. We call $\nu_{l,0}$, $\nu_{0,g}$ and $\nu_{l,g}$ the corresponding Lévy intensities and assume that they are defined as

$$\begin{aligned} \nu_{l,0}(ds, d\theta) &= c(1-z)r \frac{e^{-s}}{s} ds P_0(d\theta), & l \in \mathcal{L} \\ \nu_{0,g}(ds, d\theta) &= c(1-z)(1-r) \frac{e^{-s}}{s} ds P_0(d\theta), & g \in \mathcal{G} \\ \nu_{l,g}(ds, d\theta) &= cz \frac{e^{-s}}{s} ds P_0(d\theta), & l \in \mathcal{L}, g \in \mathcal{G}, \end{aligned}$$

where $z, r \in [0, 1]$ and P_0 is a diffuse probability measure on some space Θ . For any $l \in \mathcal{L}$ and $g \in \mathcal{G}$, we define $\tilde{\mu}_{l,g}$ as the superposition of three independent CRMs, that is

$$\tilde{\mu}_{l,g} = \mu_{l,g} + \mu_{l,0} + \mu_{0,g}. \quad (1)$$

Marginally, the $\tilde{\mu}_{l,g}$'s are identically distributed gamma CRMs with Lévy intensity $\nu(ds, dx) = ce^{-s}/s ds P_0(dx)$. While $\mu_{l,g}$ only appears in the definition of $\tilde{\mu}_{l,g}$, the components $\mu_{l,0}$ and $\mu_{0,g}$ are shared by all the $\tilde{\mu}_{l,g}$ with common l and common g , respectively, and thus induce dependence across the components of the vector $(\tilde{\mu}_{1,1}, \tilde{\mu}_{1,2}, \dots, \tilde{\mu}_{L,G-1}, \tilde{\mu}_{L,G})$. In turn, by normalizing the CRMs $\tilde{\mu}_{l,g}$, we obtain a vector $\tilde{\mathbf{p}} = (\tilde{p}_{1,1}, \tilde{p}_{1,2}, \dots, \tilde{p}_{L,G-1}, \tilde{p}_{L,G})$ of two-level GM-DDP, whose elements are identically distributed Dirichlet processes $\tilde{p}_{l,g}$, with total mass $c > 0$ and base measure P_0 . We will use the notation $\tilde{\mathbf{p}} \sim \text{GM-DDP}(c, z, r, P_0)$. Each $\tilde{p}_{l,g}$ admits the convenient decomposition

$$\tilde{p}_{l,g} = w_{l,g} p_{l,g} + (1 - w_{l,g}) v_{l,g} p_{l,0} + (1 - w_{l,g})(1 - v_{l,g}) p_{0,g},$$

where $p_{l,g}$, $p_{l,0}$ and $p_{0,g}$ are independent Dirichlet processes distributed, respectively, as the processes obtained by normalizing $\mu_{l,g}$, $\mu_{l,0}$ and $\mu_{0,g}$. For every $l \in \mathcal{L}$ and $g \in \mathcal{G}$, the random weights $w_{l,g}$ and $v_{l,g}$ are given by $w_{l,g} = \mu_{l,g}(\Theta) / (\mu_{l,g}(\Theta) + \mu_{l,0}(\Theta) + \mu_{0,g}(\Theta))$ and $v_{l,g} = \mu_{l,0}(\Theta) / (\mu_{l,0}(\Theta) + \mu_{0,g}(\Theta))$. The vectors $\mathbf{w} = (w_{1,1}, w_{1,2}, \dots, w_{L,G-1}, w_{L,G})$ and $\mathbf{v} = (v_{1,1}, v_{1,2}, \dots, v_{L,G-1}, v_{L,G})$ are marginally independent and distributed according to the $(L \times G)$ -dimensional Beta distribution defined by [10]. More specifically,

$$\begin{aligned} \mathbf{w} &\sim \text{mult-Beta}_{L \times G}(cz, \dots, cz, c(1-z)) \\ \mathbf{v} &\sim \text{mult-Beta}_{L \times G}(c(1-z)r, \dots, c(1-z)r, c(1-z)(1-r)). \end{aligned}$$

Standard properties of gamma random variables conveniently imply independence between the random probability measures $p_{l,g}$, $p_{l,0}$, $p_{0,g}$ and the random weights \mathbf{w} and \mathbf{v} .

2.2 Mixture model for partially exchangeable data

The components of $\tilde{\mathbf{p}}$ can be used as mixing random probability measures in order to define a vector of dependent mixture models. For every l, g and for any $i \in \{1, \dots, n_{l,g}\}$, let $\theta_i^{(l,g)} = (m_i^{(l,g)}, \tau_i^{(l,g)}) \in \mathbb{R} \times \mathbb{R}^+$ and $\boldsymbol{\theta}^{(l,g)} = (\theta_1^{(l,g)}, \dots, \theta_{n_{l,g}}^{(l,g)})$, and let $k(x; \boldsymbol{\theta})$ denote a univariate Gaussian kernel with parameters $\boldsymbol{\theta}$. The two-level GM-DDP mixture model for the SDSS data can be described by means of its hierarchical form as

$$\begin{aligned} (X_{i_{1,1}}^{(1,1)}, \dots, X_{i_{L,G}}^{(L,G)}) \mid \boldsymbol{\theta}^{(1,1)}, \dots, \boldsymbol{\theta}^{(L,G)} &\stackrel{\text{ind}}{\sim} \prod_{l=1}^L \prod_{g=1}^G k(x_{i_{l,g}}^{(l,g)}; \boldsymbol{\theta}_{i_{l,g}}^{(l,g)}) \\ \boldsymbol{\theta}_i^{(l,g)} \mid \tilde{\mathbf{p}} &\stackrel{\text{iid}}{\sim} \tilde{p}_{l,g} \quad l \in \mathcal{L}, g \in \mathcal{G}, i = 1, \dots, n_{l,g} \\ \tilde{\mathbf{p}} &\sim \text{GM-DDP}(c, z, r, P_0). \end{aligned}$$

It is interesting to interpret the role of the parameters z and r in modelling the SDSS data. The first one controls the weight given by the prior processes $\tilde{p}_{l,g}$ to the id-

idiosyncratic component $p_{l,g}$ and thus affects the strength of the borrowing of information across groups. The latter parameter controls the relative weight given by $\tilde{p}_{l,g}$ to the components $p_{l,0}$ and $p_{0,g}$, thus determining which feature (common luminosity or common environment) is more impactful in inducing association across groups.

3 Results

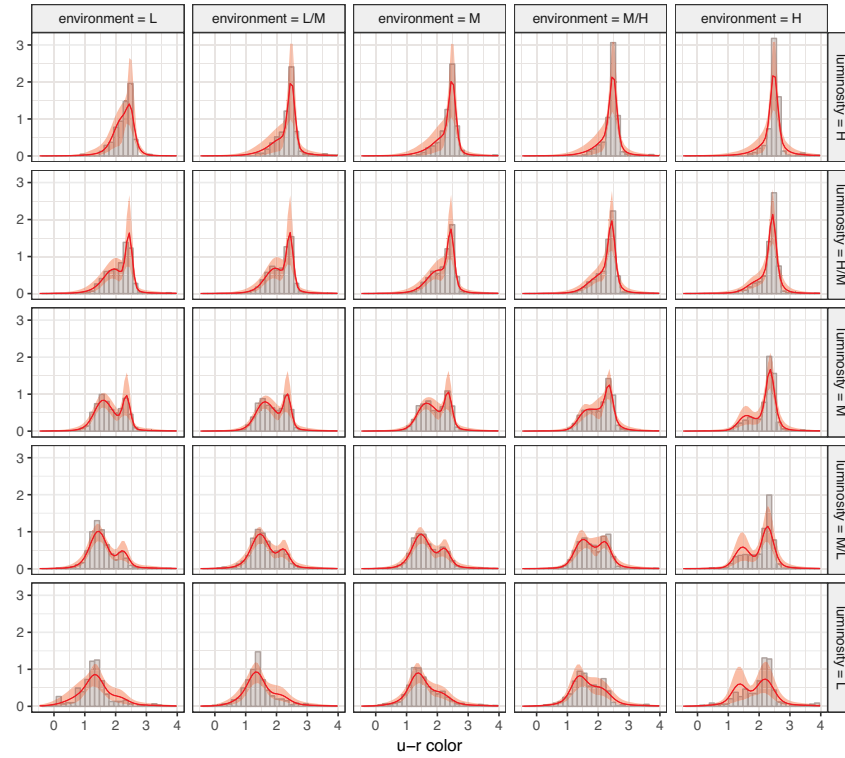


Fig. 1 Sloan Digital Sky Survey data, $u-r$ color distributions of galaxies stratified by means of co-variate information on luminosity (Z_1) and galaxy environment (Z_2). Each panel shows histogram, posterior estimated distribution (red line) and 90% posterior credible bands (red filled area). The groups correspond to *low*, *low/medium*, *medium*, *medium/high* and *high* values for luminosity and environment.

The model described in Section 2.2 is completed by setting $c = 1$ and assuming a normal-gamma base measure. More specifically, P_0 is defined so that if $(m, \tau) \sim P_0$, then $\tau \sim G(4, 1)$ and $m \mid \tau \sim N(2, 5\tau^{-1})$. The parameters z and r are both set equal to 0.5 so to formalise the idea that, a priori, idiosyncratic and common components

have the same weight and that, in turn, common luminosity and common environment are equally impactful in inducing borrowing of information across groups. The goal of our analysis consists in estimating the density function or the $u - r$ color of galaxies in each one of the 25 considered groups. To this end, the two-level GM-DDP mixture model was implemented and samples from the posterior distribution of $\tilde{\mathbf{p}}$ were obtained via a conditional Gibbs sampling algorithm, ran for 15 000 iterations, the first 5 000 of which were discarded as burn-in period. Convergence of the chain was assessed by visual investigation of the traceplots which provided no indication against it. Figure 1 displays the estimated densities, endowed with 90% posterior credible bands. A visual investigation of the posterior suggests that the borrowing of strength across strata has the effect of stressing the bimodality of the distribution of certain groups for which the histogram does not provide the same evidence. This effect can be particularly appreciated for groups with both low luminosity and high environment, such as, for example, the bottom right panel of the figure.

Acknowledgments

The authors are grateful to Michael Balogh for providing the dataset, and to Carlo Nipoti for fruitful comments on the astronomical problem.

References

1. Balogh, M. L., Baldry, I. K., Nichol, R., Miller, C., Bower, R., Glazebrook, K.: The Bimodal Galaxy Color Distribution: Dependence on Luminosity and Environment. *The Astrophysical Journal Letters* **615**.2, 101 (2004)
2. Daley, D. J. and Vere-Jones, D.: An introduction to the theory of point processes. Vol. II. Springer, New York (2008)
3. Foti, N. J., Williamson, S.: A Survey of Non-Exchangeable Priors for Bayesian Nonparametric Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **37** (2015)
4. Griffin, J. E., Kolossiaty, M., Steel, M. F.: Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society: Series B*. **75**.3, 499–529 (2013)
5. Griffiths, R., Milne, R.: A class of bivariate Poisson processes. *Journal of Multivariate Analysis*. **8**.3, 380–395 (1978)
6. Lijoi, A., Nipoti, B., Prünster, I.: Bayesian inference with dependent normalized completely random measures. *Bernoulli*. **20**.3, 1260–1291 (2014)
7. Lijoi, A., Nipoti, B., Prünster, I.: Dependent mixture models: Clustering and borrowing information. *Computational Statistics & Data Analysis*. **71**, 417–433 (2014)
8. MacEachern, S.N.: Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*. (1999)
9. MacEachern, S.N.: Dependent Dirichlet processes. Technical report. 2000.
10. Olkin, I., Liu, R.: A bivariate beta distribution. *Statistics & Probability Letters*. **62**.4, 407–412 (2003)

A case for order optimal matching: a salary gap study

Un algoritmo di matching ottimale ordinato per un studio sulle differenze salariali

Massimo Cannas

Abstract Matching algorithms are commonly used for balancing covariates in observational studies but they may give unsatisfactory results when a complete matching is not possible and matching priority should be considered. In this paper we describe an algorithm for finding an order optimal matching. An application is described through a case study concerning gender gap of ranked women and men executives.

Abstract Gli algoritmi di pair-matching sono comunemente usati per bilanciare le covariate negli studi osservativi ma possono dare risposte insoddisfacenti quando è necessario considerare anche la priorità di match, ad esempio perché non tutte le unità possono essere accoppiate. In questo lavoro descriviamo un algoritmo che trova un match ottimale in base ad un ordine di match prestabilito. L'algoritmo è illustrato tramite un caso studio nel riguardante le differenze salariali tra dirigenti di sesso diverso.

Key words: Pair-matching; incomplete matching; assignment problem.

1 Introduction

Matching algorithms are an established method for balancing covariates across two sample of units. In their simplest form, 1:1 matching, they are designed to build matched pairs where the covariate distribution is the same across pair units. *Caliper matching* refers to dropping units outside a specified matching cost (i.e. the caliper). Despite reducing the imbalance in covariate distribution the use of caliper introduces another dangerous kind of bias. The concept of “bias due to incomplete matching” has been introduced by Rosenbaum and Rubin (1985) to refer at the bias induced by an incomplete matching, i.e., one not including all treated units. The authors point

Massimo Cannas

University of Cagliari, Viale Sant'Ignazio 84, Cagliari e-mail: massimo.cannas@unica.it

at the undesirability of drops as they imply a change in the target parameter. Some tricks to help minimize the bias due to incomplete matching have been suggested in the literature on *greedy* matching. The general hint is to start the matching algorithm with “difficult to match” units in order to offer these units a bigger pool of potential controls Imbens and Rubin (2015). The idea however does not work in the context of *optimal* matching as it relies upon the sequential nature of a greedy matching algorithm.

Motivation from this work arose from a gender gap study where women executives had to be matched with similar male counterparts in a number of firms, high rank women having greater priority. In some companies the cost optimal matching resulted in high rank women sacrificed for a tiny decrease in overall cost. However, a compromise could be found between the cost and the order criteria.

In the next section we recall Gale’s result on the existence of an order optimal matchable subset. Then we describe an algorithm for finding the optimum (Section refsec:app). A gender gap study is described in Section 4.

2 Order optimal matching

Optimal matching is an established method of analysis since the seminal paper by Rosenbaum (1989). Rosenbaum (1989) briefly touches upon the problem of incomplete matching observing that “the method described [...] will always find a complete pair matching if it exists and otherwise we’ll find an optimal matching of maximum size” (Rosenbaum, 1989, p. 1027). However, the classic optimal matching may not be necessarily optimal with respect to a non uniform priority order. In this paper we propose an algorithm for finding a maximum size matching which is optimal with respect to a given matching priority. This order optimum matching may coincide with the cost optimum, being thus a doubly optimal matching. Otherwise it is possible to combine these approaches for a compromise between order and cost optimality. Let us consider a matching problem where the units to be matched are sorted in order of importance. We naturally convey matching importance using indexes, i.e. t_1 is the treated unit having highest matching priority.

Definition 1. *The set T of all treated units is matchable if different controls can be matched to each t in U .*

In general, T will not be matchable and therefore the problem becomes that of matching the most important units of T . It is not obvious that this problem has a solution since for example the matchable subset t_1, t_2 cannot be compared straightforwardly with t_2, t_3, t_4 : the former matches the two most important units; the latter matches more units. However, it has been proved by Gale (1968) that there always exists a matchable subset which is order optimal in the following sense

Definition 2. *A matchable subset t_1, \dots, t_n is said to be order optimal if, given any other matchable subset u_1, \dots, u_m , we have $m \leq n$ and $t_i \leq u_i$ for all $i \leq m$.*

In fact, it turns out that the family \mathcal{A} of all matchable subsets of T is a matroid. A proof can be found in Cook et al. (1988) but we can observe that property 1 is clearly satisfied since any subset of a matchable subset is matchable. Having proved that \mathcal{A} is a matroid the existence of an order optimal element in \mathcal{A} is an immediate consequence of the following theorem, due to D. Gale:

Theorem 1. (Gale, 1968) *Let \mathcal{A} be a family of subsets of T . If \mathcal{A} is a matroid then \mathcal{A} admits an order optimal element for any possible order.*

In the next section we show how Gale’s result can be used for an order optimal matching strategy.

2.1 A two step matching strategy

The order optimal matching implied by Theorem 1 need not to be unique since many matchings can be built on the order optimal subset. However it can be easily proved that all these matchings have the same size. Thus, in the family of maximum size matchings the subsets of order and cost optimal matchings do not necessarily intersect. If they do, their intersection is the subset of minimum cost matchings built on the optimal matchable subset of Definition 2. Thus, these matchings are, in a sense, “doubly” optimal. If the intersection is empty any maximum size matching sacrifices either the cost or the order criterion. The following scheme suggests the following mixed strategy when a complete matching is not possible:

Scheme 1. *The strategy consists in two steps:*

1. *find the matchable subset which is optimal under Definition 2;*
2. *find the minimum cost matching on this subset.*

In fact, this strategy leads to the doubly optimal matching if it exists, otherwise it finds the order optimal matching with (local) minimum cost. The latter may not be far from the global minimum, for example the matching shown in Figure 1 for the gender gap study was obtained in this way: the average matching cost was less than $0.15\sigma_{cost}$ and it is compensated by the fact that most important units have been matched (see Section 4).

2.2 Algorithm for order optimal matching

Gale’s proof of the existence of an optimal matching is not constructive.

Instead, following a suggestion of Anderson (Anderson, 1989, Ch. 3) we implement an algorithm for finding the optimum which runs in $\mathcal{O}(n^2)$. The basic idea is to greedily match treated units in decreasing order of priority trying to correct backward previous matches when the greedy matching fails.

The algorithm can be useful for statistical matching when the discrepancy matrix imposes some restrictions and this usually occurs when matching with a caliper. Otherwise the algorithm finds a trivial optimal subset of treated units. In particular i) if Hall's condition is satisfied it is always possible to find an unmatched control when greedy matching breaks down. Then the algorithm returns the whole set of treated units (or the first $|C|$ if the treated are more than the controls); ii) if Hall's condition is not satisfied some treated units are drop so we have a trivial solution where the drops are the least important treated units.

3 Statistical applications and R implementation

Order optimal matching can be implemented via the R-package OSDR. The package also contains the data used in the gender gap study described in next Section. To illustrate consider the matrix (`dist`) of the simple example discussed in previous section. In R we load package OSDR and we call function `OSDR` to implement the order optimal matching algorithm:

```
# M is a list of within-caliper control units
M1<-c("A", "B", "C")
M2<-c("A", "C")
M3<-c("B")
M4<-c("A", "C")
M5<-c("A", "D")
M  <-list(M1,M2,M3,M4,M5)
OSDR(M)

$OSDR
[1] "A" "C" "B" "0" "D"
$matched
[1] 1 2 3 5
$unmatched
[1] 4
```

The `OSDR` function gives a message each time Hall's condition is not satisfied and finally returns a vector (`opt`) showing an order optimal matching, with zero in position i meaning that t_i was drop and the subset of matched and unmatched treated units.

4 The gender gap case study

In this section we expose a case study arising from a consultancy given to a firm offering services for health resources management. The project required a fair com-

parison of several work outcomes across men and women workers and involved several firms in different sectors. The procedure for collecting information involved the use of archival data via randomized stratified sampling. Sampling strata were defined by `sex` (female=1 or male=0) and the following matching variables: `age` (years), `position` (top manager=4, middle manager/executive=3, first line manager/supervisor=2), `education` (Post-graduate=5, Graduate=4, High school=3), `contract type` (fixed term=4, permanent=3) and `seniority` (years in the position).

Matching women and men having the same job characteristics has a strong intuitive appeal and it also guarantee the separation of the design stage (where matched pairs are obtained) from the analysis stage (where the outcomes are compared). Statistical analysis based on matched pairs is commonly used in diversity management studies since the classic work of Lyness and Thompson (1997). There is evidence from gender studies literature that gender gap increases with position in the organizational hierarchy, with highest differences in the highest executive levels where pay levels and the variable part of remuneration are more variable (Lyness and Thompson, 1997; Joshi et al, 2015) so it is desirable to give matching priority to these women. However, classic optimal pair matching may drop some of the women in apical positions for a small improvement in the overall matching cost, due to caliper restrictions making impossible a complete matching or as a likely consequence of having more women than men, or both. Instead, we first restrict the dataset to the order optimal subset of women and then seek the cost optimal matching for that subset (following Scheme 1). Women executives were first ordered by (decreasing) position in the organizational hierarchy (in case of ties seniority was used to break the ties). We describe the results for firm number 6, a company showing a strong prevalence of women executives.

The order and cost optimal matchings are represented in Figure 1 with both men and women ordered by decreasing hierarchical position. It is interesting noting that the order optimal matching tends to exactly match on the priority variable while the cost optimal does not follow any matching pattern. Intuitively this is due to the sequential way the matching algorithm finds the matching compared with the global approach of the optimal matching where the problem is transformed in a classic min cost flow problem.

5 Discussion

Optimal matching is routinely used by researchers in observational studies and efficient computational implementations is now available in most statistical packages (Hansen and Olsen, 2006). Ideally, in an observational study the number of controls should be much greater than the number of treated units to make relatively easy finding a complete matched subset. However, in many situations this is not possible.

In this paper we showed that classic optimality is a particular case of order optimality corresponding to a uniform matching priority order and described an al-

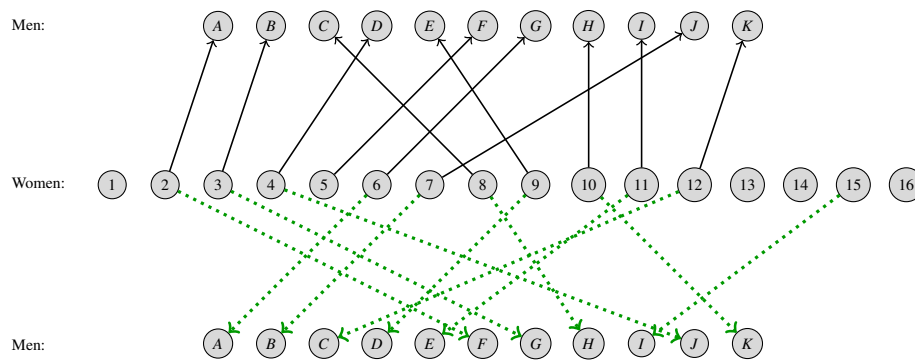


Fig. 1 Cost optimal matching (dotted lines) and order optimal matching (solid lines) for women in the gender gap study. Both matchings have the same (maximum) size but woman 7 is matched in the order optimal instead of women 15.

gorithm to efficiently find the order optimal matching. We think the algorithm can be beneficial to researchers willing to use optimal matching methods, in particular when a complete matching is not possible and the researcher would like to drop units according to a specified matching priority.

References

- Anderson I (1989) A first course in Combinatorial Mathematics. Oxford University Press
- Gale D (1968) Optimal matching in an ordered set: an application of matroid theory. *Journal of Combinatorial Theory* 4:176–180
- Hansen BB, Olsen KS (2006) Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics* 15(3):609–627
- Imbens GW, Rubin DB (2015) Causal Inference for Statistics, Social and Biomedical Sciences: an Introduction. Cambridge University Press
- Joshi A, Son J, Roh H (2015) When can women close the gap? a meta-analytic test of sex differences in performance and rewards. *Academy of Management Journal* 58(5):1516–1545, <https://doi.org/10.5465/amj.2013.0721>
- Lyness KS, Thompson DE (1997) Above the glass ceiling? a comparison of matched samples of female and male executives. *Journal of Applied Psychology* 82(3):359–375
- Rosenbaum PR (1989) Optimal matching for observational studies. *Journal of the American Statistical Association* 84(408):1024–1032
- Rosenbaum PR, Rubin DB (1985) The bias due to incomplete matching. *Biometrics* 41:106 – 116

A Prediction Method for Ordinal Consistent Partial Least Squares

Un Metodo di Previsione per l'Algoritmo Ordinal Consistent Partial Least Squares

Gabriele Cantaluppi and Florian Schuberth

Abstract We present a prediction method for ordinal partial least squares and ordinal consistent partial least squares. Both are variance-based estimators similar to partial least squares path modeling and consistent partial least squares, but take into account the non-metric scale of ordinal categorical indicators.

Abstract *Si presenta un metodo di previsione per gli algoritmi ordinal partial least squares e ordinal consistent partial least squares, stimatori variance based simili a partial least squares e consistent partial least squares che tengono conto della natura non metrica delle scale in presenza di indicatori categorici di tipo ordinale.*

Key words: ordinal consistent partial least squares, partial least squares, out-of-sample prediction

1 Introduction

Partial least squares path modeling (PLS) is a variance-based estimator in structural equation modeling that first creates linear combinations of observable indicators as stand-ins for the theoretical concepts¹ and subsequently estimates the model param-

Gabriele Cantaluppi

Faculty of Economics, Department of Statistical Science, Università Cattolica del Sacro Cuore, Largo Gemelli, 1, 20123 Milan, Italy, Tel.: +390272342492, e-mail: gabriele.cantaluppi@unicatt.it

Florian Schuberth

Faculty of Engineering Technology, University of Twente, Horst Complex W246, P.O. Box 217, 7500 AE Enschede, The Netherlands, Tel.: +31534897846, e-mail: f.schuberth@utwente.nl

¹ A theoretical concept is part of the theoretical model, while a construct represents its counterpart in the statistical model [11].

eters [16]. In addition to the usefulness of PLS in confirmatory research [6], in recent years, PLS's predictive capacities has gained increasingly more attention [1, 4, 10].

However, all major developments and extensions of PLS for predictive modeling assume metric observable indicators, which are often not available in empirical research as the dataset was, e.g., collected by questionnaires, often resulting in ordinal categorical indicators. To address this shortcoming, several approaches have been developed to deal with the non-metric nature of the indicators in the context of PLS such as non-metric partial least squares [12], partial maximum likelihood partial least squares [7], ordinal partial least squares [2, 3], and ordinal consistent partial least squares (OrdPLSc) [13, 14].

OrdPLS and OrdPLSc (OrdPLS(c)) were introduced in the context of PLS to address with ordinal categorical indicators in a psychometric way. Both approaches assume that the ordinal categorical indicators are crude measures of underlying latent variables and therefore use the polychoric correlation matrix as input to the PLS algorithm. While OrdPLS only consistently estimates structural models containing composites, OrdPLSc is capable of consistently estimating structural models containing both composites and common factors by applying a correction for attenuation as consistent partial least squares (PLSc) [5]. Although OrdPLS(c) can be used for model estimation, they cannot currently be employed for predictive modeling because of their assumption about the nature of ordinal categorical indicators.

Against this background, we contribute a prediction method for OrdPLS(c) that allows us to conduct out-of-sample predictions using OrdPLS(c). The remainder of the paper is structured as follows: Section 2 presents the developed prediction method for the OrdPLS(c) estimators. In Section 3, we conduct a Monte Carlo simulation to evaluate the performance of the proposed prediction method and to compare it to the method usually adopted in context of PLS and PLSc (PLS(c)) when ordinal categorical indicators are addressed. Moreover, we present the employed evaluation measures. In Section 4, we present the results of our simulation and a brief discussion.

2 Prediction Method for OrdPLS and OrdPLSc

OrdPLS(c) work similarly to PLS(c); however, they can address ordinal categorical indicators in a psychometric way. In doing so, it is assumed that an ordinal categorical indicator x is the outcome of a polytomized standard normally distributed latent random variable x^* :

$$x = m \quad \text{if} \quad \tau_{m-1} \leq x^* < \tau_m \quad m = 1, \dots, M, \quad (1)$$

where the threshold parameters τ_0, \dots, τ_M determine the observed category. The first and the last thresholds are fixed: $\tau_0 = -\infty$ and $\tau_M = \infty$. Moreover, the thresholds are assumed to be strictly increasing: $\tau_0 < \tau_1 < \dots < \tau_M$. In case of more than one categorical indicator, it is assumed that the categorical indicators \mathbf{x} are the outcome

of categorized underlying multivariate standard normally distributed latent random variables \mathbf{x}^* . Consequently, the observations of the indicators \mathbf{x}_j belonging to construct j , which are stored in the matrix \mathbf{X}_j , are the outcome of columnwise transformations (as expressed by Equation 1) of the observations of the underlying multivariate normally distributed random variables that are stacked in the matrix \mathbf{X}_j^* :

$$\mathbf{X}_j^* \rightarrow \mathbf{X}_j. \quad (2)$$

In contrast to PLS(c), which use the Pearson correlations among the observed categorical indicators, OrdPLS(c) employ the polychoric correlation as the input for the PLS algorithm. Consequently, no construct scores, i.e., factor and composites scores, are directly available in OrdPLS(c). However, these scores can be obtained by applying the so-called ‘mean’, ‘median’, and ‘mode’ methods [2, 13].

In the following sections, we propose a prediction method for OrdPLS(c) to obtain predictions at the item level of the endogenous constructs, starting from the categorical indicators of the exogenous constructs. Our starting point is a new set of observations $\mathbf{X}_{j,exog}^o$, $j = 1, \dots, J_{exog}$ of the ordinal categorical indicators $\mathbf{x}_{j,exog}$ belonging to the exogenous constructs in the structural model. This set of new observations $\mathbf{X}_{j,exog}^o$ can be assumed to be the columnwise transformations, as stated by Equation 1, of observations of a multivariate *truncated* normally distributed dataset $\mathbf{X}_{j,exogTRUNC}^*$ on the ordinal scale:

$$\mathbf{X}_{j,exogTRUNC}^*(continuous) \rightarrow \mathbf{X}_{j,exog}^o. \quad (3)$$

The dataset $\mathbf{X}_{j,exogTRUNC}^*$ is standardized and has the same correlation matrix as the polychoric correlation of exogenous manifest variables of the original dataset. The domain of $\mathbf{X}_{j,exogTRUNC}^*$ is defined by the same thresholds (τ_{j-1}, τ_j) obtained when computing the polychoric correlation matrix on the original dataset.

- Step 1: Calculate the construct scores of the exogenous constructs. In PLS(c), construct scores for construct j are defined as a linear combination of the observable indicators \mathbf{x}_j belonging to this construct. To obtain construct scores in OrdPLS(c), the unknown values of the continuous latent variables underlying the categorical indicators need to be aggregated. To address this issue, the construct scores of the exogenous constructs are defined as linear combinations of multivariate *truncated* normally distributed random variables $\mathbf{x}_{j,exogTRUNC}^*$, which are continuous with domain (τ_{j-1}, τ_j) . Consequently, the distribution of the construct scores is a linear combination of truncated normally distributed random variables of the type displayed in Equation 3 with weights estimated by OrdPLS(c) based on the original dataset. The distribution of the construct scores has no explicit simple form but can be approximated by simulation. In doing so, a sufficient number of drawings, e.g., $n_{pred} = 100$, from each truncated multivariate normal distribution² $\mathbf{X}_{j,exogTRUNC}^*$, corresponding to $\mathbf{x}_{j,exog}^o$, is used to calculate the linear combinations. As a result, we obtain n_{pred} predictions of possible sets of construct

² Possibly adjusting relationship displayed in Equation 1 in order to take into account categories that were not observed in the training set.

scores for each subject, consistent with the observed categories, that we denote as follows:

$$\hat{\eta}_{j,exog}^{o,p} = \mathbf{X}_{j,exog}^* \text{TRUNC} \hat{\mathbf{w}}_j, \quad j = 1, \dots, J_{exog}, \quad p = 1, \dots, n_{pred}. \quad (4)$$

- Step 2: Predict the scores of the endogenous constructs in accordance with the structural model: $\hat{\eta}_{endog}^{o,p} = (\mathbf{I} - \hat{\mathbf{B}})^{-1} \hat{\mathbf{F}} \hat{\eta}_{exog}^{o,p}$, where $\hat{\mathbf{B}}$ and $\hat{\mathbf{F}}$ contain path coefficient estimates based on the original dataset. As an outcome, we obtain n_{pred} predicted values for the endogenous constructs.
- Step 3: Predict the continuous latent variables underlying the categorical indicators belonging to the endogenous constructs. A set of n_{pred} predictions for the indicators belonging to the j -th endogenous constructs, $\hat{\mathbf{x}}_{j,endog}^*$ with $j = J_{exog} + 1, \dots, J$, can be obtained by multiplying the prediction of the endogenous construct $\hat{\eta}_{j,endog}^{o,p}$ with the corresponding loading estimates based on the original dataset.
- Step 4: Predict the values of the observable ordinal categorical indicators belonging to the endogenous constructs. To obtain predictions on ordinal scales, the n_{pred} (continuous) predictions of the components of $\hat{\mathbf{x}}_{j,endog}^*$ can be summarized by their mean or their median and subsequently transformed according to Equation 2 by using the threshold parameter estimates based on the original dataset. Moreover, a *mode* estimate of the category can be obtained from the n_{pred} estimates of the components of $\hat{\mathbf{x}}_{j,endog}^*$ by considering the maximum empirical density on the intervals defined by the thresholds in Equation 2.

3 Monte Carlo Simulation

To evaluate the proposed prediction method for OrdPLS(c) and compare its performance to more commonly used approaches in the context of PLS and PLS_c, i.e., treat the categorical as continuous and use rounded predictions, we conducted a Monte Carlo simulation. Accordingly, we focused on out-of-sample casewise operative predictions [15], i.e., new observations of the indicators of the exogenous constructs are used to predict the values of the indicators of the endogenous constructs while these new observations were not used for the estimation of the model parameters. For the simulation design, we chose the exact same setting as suggested by [14] for the assessment of OrdPLS(c): two population models (a model with three common factors and a model with one common factor and two composites), different numbers of consecutive categories (2, 3, 4, 5, and, 7 categories), different ordinal categorical indicator distributions (symmetric, moderate asymmetric, extreme asymmetric, alternating moderate asymmetric, and alternating extreme asymmetric), and different sample sizes (250 and 500 observations). For each condition, we draw 1,000 multivariate standard normally distributed training sets that are used for the model estimation and corresponding test datasets containing 1,000 observations for the indicators of the exogenous constructs. The continuous indicators were categorized according to Equation 1. To approximate the distribution of the scores of

the exogenous constructs, we set n_{pred} to 100. Each condition was estimated by OrdPLSc, PLSc, OrdPLS, and PLS, and inadmissible estimates were removed.

To assess the predictive performance, we considered the mean squared error (MSE , [8]) and the proportion of concordance, both computed based on the test dataset. The MSE is defined as $MSE_k = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{x}_{ij} - x_{ij})^2$, $k = 1, 2, \dots, 1000$, where x_{ij} represents the actual observed category of the j -th indicator for the i -th observation in the test data set and \hat{x}_{ij} is the corresponding predicted category. As we perform 1,000 simulation runs per condition, 1,000 estimates of MSE are available. Usually, the MSE is only used for metric data. Here, we consider it as an evaluation measure assigning a penalty of 0 to exact concordances between actual and predicted categories, a penalty of 1 if category $h - 1$ or $h + 1$ is predicted for the observed category h , a penalty of 4 if category $h - 2$ or $h + 2$ is predicted for category h , and so on. Furthermore, we define *concordance* as the proportion of correct predictions over all observations in each test data set $conc_k = \frac{1}{1000} \sum_{i=1}^{1000} I(\hat{x}_{ij}, x_{ij})$, $k = 1, 2, \dots, 1000$, where $I(\cdot)$ is the indicator function that takes value 1 when $\hat{x}_{ij} = x_{ij}$ and 0 otherwise. We deliberately did not analyze the residuals of the continuous variables underlying observed predicted categorical variables (see e.g., [9]) since we want to compare the performance of the OrdPLS(c) to traditional methods that treat categorical indicators as continuous.

We further studied the relative performance in terms of the same evaluation measures when comparing the proposed procedures first with PLS and then with PLSc. PLSc and OrdPLSc are consistent estimators in the framework of continuous and ordinal indicators, respectively [5, 13, 14]. Here, we examined their predictive behavior when ordinal categorical variables are used instead of continuous ones.

4 Results and Discussion

In general, it should be noted that both the MSE and the concordance evaluation measures led to similar results with negligible differences among estimation methods for the symmetric, moderate asymmetric, and alternating moderate asymmetric threshold parameter distributions for both population models.

In all situations, the MSE increases and the *concordance* decreases when the number of scale points increases. Both the MSE and *concordance* evaluation measure values are practically equivalent for OrdPLS(c) and rounded PLS(c) with regard to endogenous manifest indicators. The worst performance of rounded PLS(c) and the highest difference between OrdPLS(c) and PLS(c) is obtained in the case of extreme asymmetric and alternating extreme asymmetric threshold parameter distributions. Namely, in the presence of (alternating) extreme asymmetric distributions, rounded PLS(c) performed slightly better than other predictor methods in terms of the MSE metric, but performed worse in terms of the *concordance* metric. The largest difference in *concordance* measures between rounded PLS(c) and OrdPLS(c) is present for (alternating) extreme distributions with 4-point scales and decreases when the number of points in the scale increases. In all these cases, Or-

dPLS(c) prediction methods perform better than rounded PLS(c) according to the *concordance* metric. The negligible difference in *MSE* denotes a larger spread between predicted and actual values for OrdPLS(c) prediction methods, though they have a larger proportion of correct predictions. It is emphasized that 4-point scales are relevant for empirical research, as they are encountered in questionnaires, e.g., when answers are coded as *very bad*, *bad*, *good*, and *very good*.

We can conclude that the prediction method for OrdPLS(c) appears to produce more reliable predictions than rounded PLS(c) methods in the case of (alternating) extreme distribution setting.

References

1. Becker, J.M., Rai, A., Rigdon, E.: Predictive Validity and Formative Measurement in Structural Equation Modeling: Embracing Practical Relevance. In: Proceedings of the International Conference on Information Systems, pp. 119 (2013)
2. Cantaluppi, G.: A Partial Least Squares Algorithm Handling Ordinal Variables also in Presence of a Small Number of Categories. arXiv preprint arXiv:1212.5049 (2012)
3. Cantaluppi, G., Boari, G.: A Partial Least Squares Algorithm Handling Ordinal Variables. In H. Abdi *et al.* (eds.) The multiple facets of partial least squares and related methods: PLS, Paris, 2014, pp. 295-306. Springer (2016)
4. Cepeda Carrión, G., Henseler, J., Ringle, C.M., Roldán, J.L.: Prediction-oriented modeling in business research by means of PLS path modeling: Introduction to a JBR special section. *J Bus Res* **69**(10), 4545-4551 (2016)
5. Dijkstra, T.K., Henseler, J.: Consistent Partial Least Squares Path Modeling. *MIS Quart* **39**(2), 297-316 (2015)
6. Henseler, J., Hubona, G., Ray, P.A.: Using PLS path modeling in new technology research: Updated guidelines. *Ind Manage Data Syst* **116**(1), 2-20 (2016)
7. Jakobowicz, E., Derquenne, C.: A modified PLS path modeling algorithm handling reflective categorical variables and a new model building strategy. *Comput Stat Data An* **51**(8), 3666-3678 (2007)
8. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning. with Applications in R. Springer (2017)
9. Liu, D., Zhang, H.: Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach. *J Am Stat Assoc* **113**, 845-854 (2018)
10. Rigdon, E.E.: Rethinking Partial Least Squares Path Modeling: In Praise of Simple Methods. *Long Range Plann* **45**(5), 341-358 (2012)
11. Rigdon, E.E., Becker, J.M., Sarstedt, M.: Factor indeterminacy as metrological uncertainty: Implications for advancing psychological measurement. *Multivar Behav Res*, 1-15 (2019)
12. Russolillo, G.: Non-metric Partial Least Squares. *Electron J Stat* **6**, 1641-1669 (2012)
13. Schuberth, F., Cantaluppi, G.: Ordinal Consistent Partial Least Squares. In: H. Latan, R. Noonan (eds.) Partial Least Squares Path Modeling. Basic Concepts, Methodological Issues and Applications, pp. 109-150. Springer (2017)
14. Schuberth, F., Henseler, J., Dijkstra, T.K.: Partial least squares path modeling using ordinal categorical indicators. *Qual Quant* **52**, 9-35 (2018)
15. Shmueli, G., Ray, S., Estrada, J.M.V., Chatla, S.B.: The elephant in the room: Predictive performance of PLS models. *J Bus Res* **69**(10), 4552-4564 (2016)
16. Wold, H.: Path Models with Latent Variables: The NIPALS Approach. In: H. Blalock *et al.* (eds.) Quantitative Sociology, International Perspectives on Mathematical and Statistical Modeling, chap. 11, pp. 307-357. Academic Press (1975)

Functional control charts for monitoring ship operating conditions and CO₂ emissions based on scalar-on-function linear model

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

Christian Capezza, Antonio Lepore, Alessandra Menafoglio, Biagio Palumbo, and Simone Vantini

Abstract Shipping companies are enforced to respond to the recent air pollution monitoring regulations implemented worldwide and to install on their fleets multi-sensor systems that automatically record data during each voyage. However, the use of this massive amount of observational data streams is still an open challenge for shipping operators. Several additional factors can in fact dramatically affect CO₂ emissions and are function of time. We propose functional control charts to (i) monitor ship operating condition profiles at each voyage in order to identify if and when potential out-of-control condition may have occurred; and to (ii) predict the total CO₂ emissions through a scalar-on-function regression model. The proposed control charts are illustrated on real data from a roll-on/roll-off passenger cruise ship.

Abstract Le normative adottate negli ultimi anni a livello internazionale sul monitoraggio delle emissioni inquinanti impongono alle compagnie armatoriali di installare a bordo delle proprie navi sistemi automatici di acquisizione di dati di navigazione. In tale ottica, il corretto utilizzo di questa grande mole di dati raccolti rappresenta pertanto un'opportunità strategica per gli armatori. Poiché le emissioni inquinanti di una nave sono potenzialmente influenzate da molti fattori variabili nel tempo, in questo lavoro proponiamo un approccio innovativo per il monitoraggio delle emissioni mediante carte di controllo funzionali per (i) il monitoraggio delle condizioni operative della nave finalizzato all'individuazione di potenziali condizioni anomale e (ii) la previsione delle emissioni totali di CO₂ attraverso un modello di regressione funzionale con risposta scalare. L'approccio proposto viene

Christian Capezza, Antonio Lepore, and Biagio Palumbo
Department of Industrial Engineering, University of Naples Federico II, P.le V. Tecchio 80, 80125 Naples, Italy, e-mail: christian.capezza@unina.it; antonio.lepore@unina.it; biagio.palumbo@unina.it

Alessandra Menafoglio and Simone Vantini
MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy, e-mail: alessandra.menafoglio@polimi.it; simone.vantini@polimi.it

applicato a un caso reale mediante dati raccolti a bordo di una nave da carico e passeggeri.

Key words: Functional data analysis, Functional control chart, Multivariate functional principal component analysis

1 Introduction

Monitoring CO₂ emissions from maritime transportation has become in recent years a compelling task in view of the extensive air pollution programs implemented by the Marine Environment Protection Committee of the International Maritime Organization [10, 11, 12, 13, 16, 14], and applied at European level by the EU regulation 2015/757 [9] that came into force from January 2018. Modern multi-sensor systems are nowadays affordably installed on-board and facilitate streaming of ship's observational data along time during each voyage, bypassing human intervention. It is known from naval literature that several additional factors may have dramatic effect on vessel performance and thus CO₂ emissions, e.g., ship type, draught, speed, acceleration, encounter angle, wind regime, sea state [1]. However, most of the approaches presented involve only summary statistics of the complete signals over each voyage.

Within this background, functional control charts are proposed to monitor ship operating conditions and total CO₂ emissions at each voyage by accounting for the influence of the complete profiles of those additional factors (i.e., *operating condition profiles*). The approach is illustrated by means of a real-case study concerning CO₂ emission data from a roll-on/roll-off passenger (Ro-Pax) cruise ship owned by the Grimaldi Group. The working pipeline is based on the two following main steps: (i) monitoring of the operating condition profiles from the beginning to the end of each voyage in order to identify potential out-of-control (OC) conditions; (ii) prediction—for each Phase II voyage estimated in normal condition—of the total CO₂ emissions from the ship observational data recorded during the voyage. The step (i) is based on the Hotelling T^2 (e.g., [7, 8]) and squared prediction error (*SPE*) control charts for the multivariate functional principal component scores pertaining to the operating condition functions. The step (ii) is based on a scalar-on-function regression model (e.g., [4]) with the total CO₂ emissions playing the role of the response and operating condition profiles being the covariates.

2 Methodology

In this communication, we illustrate the methodology we developed in [3], to which we refer for further details.

2.1 Scalar-on-function linear model

Observations of functional covariates can be embedded in the Hilbert space \mathbb{H} of P -dimensional vectors whose components are functions in the space $L^2(\mathcal{T})$, with $\mathcal{T} = (0, 1)$. Given two elements of \mathbb{H} , $f = (f_1, \dots, f_P)$ and $g = (g_1, \dots, g_P)$, the inner product of \mathbb{H} can be defined as $\langle f, g \rangle_{\mathbb{H}} = \sum_{p=1}^P \langle f_p, g_p \rangle$, where $\langle f_p, g_p \rangle = \int_{\mathcal{T}} f_p(t)g_p(t)dt$ is the inner product of $L^2(\mathcal{T})$. The induced norm of \mathbb{H} is $\|f\|_{\mathbb{H}} = \langle f, f \rangle_{\mathbb{H}}$. Let us denote with $\mathbf{X} = (X_1, \dots, X_P)$ a random element that takes values in \mathbb{H} . Without loss of generality, we assume $E(X_p(t)) = 0$ and $\text{Var}(X_p(t)) = 1 \forall t \in (0, 1)$ and $\forall p = 1, \dots, P$. Let us denote with y the scalar response variable and consider a random sample $\{(\mathbf{X}_i, y_i)\}_{i=1, \dots, n}$ from (\mathbf{X}, y) to be used as the Phase I reference dataset. The scalar-on-function regression model is

$$y_i = \beta_0 + \langle \mathbf{X}_i, \boldsymbol{\beta} \rangle_{\mathbb{H}} + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{H}$ are the coefficient to be estimated and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. error terms, assumed distributed as $N(0, \sigma^2)$ and independent of the functional covariates.

2.2 Phase I Model Estimation

The coefficients β_0 and $\boldsymbol{\beta}$ in Eq.(1) can be estimated using least-squares. However, the problem is not well-posed, then it can be approached through multivariate functional principal component analysis [4, 5] by considering a finite-dimensional approximation of $\mathbf{X}(t)$ obtained through the Karhunen-Loève expansion. The choice of the multivariate functional principal components to retain in the model can be carried out by maximizing the proportion of the total variability explained by the principal components, i.e. by retaining the first M components [6]. However, if one is interested in prediction of the scalar response, there are some components that may have small predictive ability. A parsimonious choice may be discarding all components whose variance is less than a threshold and result not significant for the regression on the scalar response, on the basis of some index on the predictive performance, e.g. using cross-validation.

2.3 Phase II Monitoring

We propose three monitoring statistics for Phase II monitoring. Suppose that a new observation $(\mathbf{X}^{new}, y^{new})$ is available from (\mathbf{X}, y) .

The Hotelling T^2 statistic monitors the components retained to estimate the scalar-on-function regression model. The squared prediction error SPE statistic looks at the squared norm of the residual function obtained by substituting \mathbf{X}^{new} with its M -dimensional approximation based on the multivariate functional principal

components. The scalar response can also be monitored itself through the response prediction error due to the regression. The latter can be regarded as the natural extension of the regression control chart known in the SPC literature firstly introduced by [15]. Nonparametric control limits of T^2 and SPE can be set on the basis of the empirical distribution, while parametric control limits of the prediction error control chart are based on the normal distribution of the response. The Bonferroni correction can be utilized to guarantee that the type-I family-wise error rate is not larger than a desired α . If at least one of the Hotelling T^2 and SPE statistics is out of the control limits, it can be decomposed as the sum of single contributions of each functional covariate. This allows identifying which variables are responsible for unusual behaviors. When an out of control is issued by the response prediction error control chart, possible causes have to be investigated outside the set of variables included in the model as functional covariates.

3 A real-case study

Data collected from the multi-sensor system installed on board of a Ro-Pax cruise ship owned by the Italian shipping company Grimaldi Group are used to build the functional covariates and illustrate the proposed method. Each observation refers to a specific voyage at given route and direction. The name of the ship, route, and voyage dates are omitted for confidentiality reasons. The response variable is the total CO₂ emissions in navigation phase per each voyage, measured in *tons*. Nine variables used as functional covariates in the scalar-on-function model, namely the cumulative sailing time, speed over ground, acceleration, power difference between port and starboard propeller shafts, distance from the nominal route, longitudinal and transverse wind components, air temperature of the main engines, trim. Information about the variables can be found in [2]. Since traveling time could differ over voyages, functional data are registered by using the fraction of distance travelled over the voyage as the common domain $\mathcal{T} = (0, 1)$.

The proposed functional control charts are demonstrated to be a powerful tool for supporting prognostic of faults, i.e. indicating the predictors that may have determined an OC on the operating conditions. Moreover, the proposed procedure is shown to well predict and monitor ship CO₂ emissions and to signal if and when an anomaly may have occurred caused by factors that are plausibly external to those collected into operating conditions.

In the talk, we illustrate relevant case studies that show the effectiveness of the proposed procedure. Moreover we show how to employ them for real-time monitoring, i.e. issue an OC signal before a voyage is complete and functions are only partially observed.

References

1. Bialystocki, N., Konovessis, D.: On the estimation of ship's fuel consumption and speed curve: A statistical approach. *Journal of Ocean Engineering and Science* **1**(2), 157–166 (2016)
2. Bocchetti, D., Lepore, A., Palumbo, B., Vitiello, L.: A statistical approach to ship fuel consumption monitoring. *Journal of Ship Research* **59**(3), 162–171 (2015)
3. Capezza, C., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Control charts for monitoring ship operating conditions and CO₂ emissions based on scalar-on-function regression. MOX-Report 12/2019 (2019)
4. Chiou, J.M., Chen, Y.T., Yang, Y.F.: Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica* pp. 1571–1596 (2014)
5. Chiou, J.M., Yang, Y.F., Chen, Y.T.: Multivariate functional linear regression and prediction. *Journal of Multivariate Analysis* **146**, 301–312 (2016)
6. Chiou, J.M., Zhang, Y.C., Chen, W.H., Chang, C.W.: A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transportmetrica B: Transport Dynamics* **2**(2), 106–129 (2014)
7. Colosimo, B.M., Pacella, M.: On the use of principal component analysis to identify systematic patterns in roundness profiles. *Quality and reliability engineering international* **23**(6), 707–725 (2007)
8. Colosimo, B.M., Pacella, M.: A comparison study of control charts for statistical monitoring of functional data. *International Journal of Production Research* **48**(6), 1575–1601 (2010)
9. Council of European Union: Regulation (EU) 2015/757 of the European Parliament and of the Council of 29 April 2015 on the monitoring, reporting and verification of carbon dioxide emissions from maritime transport, and amending directive 2009/16/EC (2015)
10. IMO: Air pollution and greenhouse gas (ghg) emissions from international shipping, marpol annex 6. london, u.k. (2012)
11. IMO: Guidelines for the development of a ship energy efficiency management plan (seemp), mepc.213(63) annex 9. london, u.k. (2012)
12. IMO: Guidelines on the method of calculation of the attained energyefficiency design index (eedi) for new ships, mepc.212 annex 8.london, u.k. (2012)
13. IMO: 2014 guidelines on survey and certification of the energy efficiency design index (eedi), london, u.k. (2014)
14. ISO: ISO 19030-1, 2, 3 Ships and marine technology—Measurement of changes in hull and propeller performance (Part 1: General principles; Part 2: Default method; Part 3: Alternative methods) (2016)
15. Mandel, B.: The regression control chart. *Journal of Quality Technology* **1**(1), 1–9 (1969)
16. Smith, T., Jalkanen, J., Anderson, B., Corbett, J., Faber, J., Hanayama, S., O'Keeffe, E., Parker, S., Johansson, L., Aldous, L., Raucci, C., Traut, M., Ettinger, S., Nelissen, D., Lee, D., Ng, S., Agrawal, A., Winebrake, J., Hoen, M., Chesworth, S., Pandey, A.: Third imo ghg study 2014. International Maritime Organization (IMO), London, UK (2015)

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-

Title Suppressed Due to Excessive Length

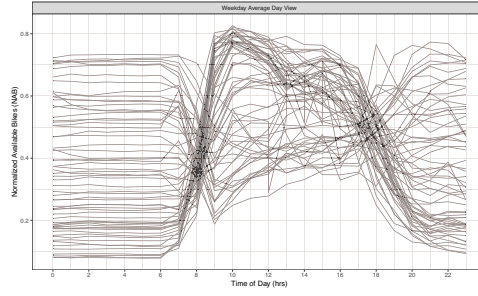


Fig. 1 Average weekday Normalized Available Bikes (number of bikes / total number of slots in the station) for the BikeMi stations in Milan central area.

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

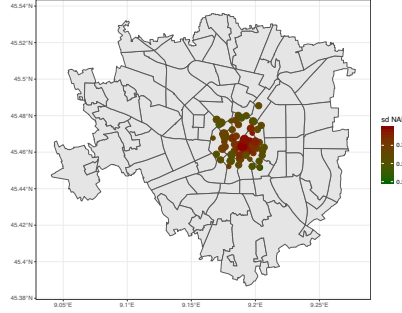


Fig. 2 Location of the BikeMi stations considered in the analysis. Dots size denotes the station total number of slots, the color scaling indicates the standard deviation of normalized bikes availability averaged per weekday.

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	Cairolì	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.

Public support for an EU-wide social benefit scheme: evidence from Round 8 of the European Social Survey (ESS)

Sostegno pubblico a un sistema di prestazioni sociali a livello dell'Unione Europea: i risultati del Round 8 della European Social Survey (ESS)

Paolo Emilio Cardone¹

Abstract The ESS Round 8 module (fielded in 2016/17) - Welfare Attitudes in a Changing Europe: Solidarities under Pressure - makes it possible to shed scientific light on welfare debates. The inclusion of the Welfare Attitudes in Europe module during Round 8 of the ESS, first of all allowed attitudes towards these services to be assessed in 23 countries but also it addresses new solidarity questions as this new module also includes some questions fielded for the first time – most notably items assessing the introduction of a universal basic income (UBI) scheme and the implementation of a European Union-wide social benefit scheme. Using logistic regression model it is possible to estimate the different attitudes among countries for an EU-wide social benefit scheme more accurately.

Abstract Il presente contributo intende focalizzare l'attenzione sul sostegno pubblico a un sistema di prestazioni sociali a livello dell'Unione Europea. L'analisi è stata effettuata utilizzando i dati della European Social Survey (ESS), una ricerca quantitativa condotta dall'Inapp (Istituto nazionale per l'Analisi delle politiche pubbliche) per l'Italia. In particolare è stata analizzata la sezione “welfare” in maniera comparativa tra i 23 Paesi Europei coinvolti nell'indagine. Utilizzando un modello di regressione logistica, è stato possibile stimare più accuratamente i fenomeni che favoriscono o meno la percezione dei cittadini verso politiche sociali condotte e gestite a livello europeo. Se, in Europa, la mancanza di un sistema europeo di welfare è avvertita come un'inefficienza che mina il consenso sull'intera costruzione europea, l'opinione in maggioranza favorevole degli intervistati all'introduzione di “un sistema di prestazioni sociali a livello dell'Unione europea per tutte le persone povere” può essere considerata una chiara indicazione di direzione del cambiamento atteso.

¹ Paolo Emilio Cardone, INAPP-Statistical Office; email: p.cardone@inapp.org

Key words: European Social Survey; Social Benefit Scheme; Welfare Attitudes; Logistic Model; Stata.

1 Introduction

An important aspect of most democratic societies is a welfare state – government funded services that offer financial protection to its citizens, paid for by taxes. This can encompass a whole plethora of services including healthcare provision, unemployment benefits, housing costs and pensions.

In the past decades, the extended European-style welfare state became substantially challenged due to a number of major economic, social and political developments. Longer-term challenges have been exacerbated by the shock of the banking crisis in 2008, which was quickly followed by an economic recession in 2009, and a longer-lasting fiscal and debt crisis in many European states.

Many countries have experienced government-imposed austerity measures since then, and many areas of public expenditure have been stagnant, scaled back or cut completely. We can now assess whether financial restrictions on the welfare state in many countries have changed public attitudes towards it.

In many welfare states, the challenges posed by the nearly universal trends of growing inequality, migration, ageing, globalisation and digitalization of work have been further aggravated by the recent economic crisis. At the same time, these trends put the sustainability of social policies under pressure and thus bring back to the political agenda discussions about policy reforms.

Finally, there is an ongoing European Union (EU) debate, ignited substantially by the unequal degree to which the economic crisis has hit the different countries in Europe. It regards the solidarity between Europeans, addressing the question of whether a redistribution of welfare from richer to poorer Europeans would be necessary to create cross-European social cohesion, and would be politically and economically feasible.

The ESS Round 8 module (fielded in 2016/17) - Welfare Attitudes in a Changing Europe: Solidarities under Pressure - makes it possible to shed scientific light on these debates.

The inclusion of the Welfare Attitudes in Europe module during Round 8 of the ESS, first of all allowed attitudes towards these services to be assessed in 23 countries but also it addresses new solidarity questions as this new module also includes some questions fielded for the first time – most notably items assessing the introduction of a universal basic income (UBI) scheme and the implementation of a European Union-wide social benefit scheme. Using logistic regression model it is possible to estimate the different attitudes among countries for an EU-wide social benefit scheme more accurately.

2 Data and methods

In order to achieve this goal, the analysis is carried out using microdata from the quantitative research “*European Social Survey*” (ESS)², provided by Inapp³.

The possibility to track, capture and investigate individuals’ behaviours, values, beliefs and attitudes over time and across space has become increasingly relevant for the scholarly understanding of a rapidly changing social world.

For nearly twenty years now, the European Social Survey (ESS) has aided researchers in this task. The ESS is an academically driven cross--- national survey that has been conducted in over thirty countries across Europe since its establishment in 2001.

Every two years, face--- to--- face interviews are conducted with newly selected, cross--- sectional samples investigating the attitudes, beliefs and behaviour patterns of diverse populations. The core and rotating modules that form the backbone of the ESS questionnaires have addressed multiple topics, including attitudes toward the media, social trust, politics, democracy and citizen involvement; subjective well--- being and human values; attitudes towards immigration; family, work and well--- being, the timing of life and gender roles; economic morality, welfare attitudes and justice; public attitudes toward climate change. Italy has participated in the ESS on four occasions: in rounds 1, 2 and 6, and in the recent round 8, collected between 2016 and 2017 and released in May 2018.

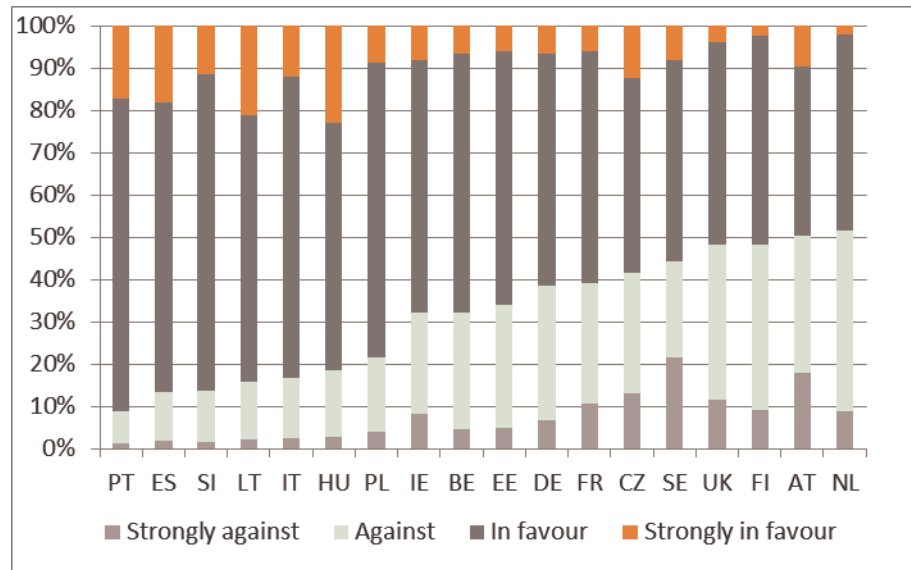
The ESS Round 8 module (fielded in 2016/17) - Welfare Attitudes in a Changing Europe: Solidarities under Pressure - makes it possible to shed scientific light on these debates.

As shown in table 1, Portugal is the most European country clearly in favour of a public support for an EU-wide social benefit scheme (in favour and strongly in favour are more than 90%) followed by Spain, Slovenia, Lithuania, Italy and Hungary with a percentage between 80 and 90%. On the other side, The Netherlands are the most opposed country with less than 50% in favour. This is why we have chosen this country as reference category in the logistic model showed below.

Table 1: *Public support for an EU-wide social benefit scheme*

² For more details: <https://inapp.org/it/dati/ESS>

³ National Institute for Public Policy Analysis, former ISFOL - National Research Institute for Vocational Education and Training Employment, that changed its company name in INAPP (*Istituto Nazionale per l'Analisi delle Politiche Pubbliche – Public Policy Innovation*) on December 1st 2016 (www.inapp.org)



Source: own elaboration on ESS data

Legend: PT=Portugal; ES=Spain; SI=Slovenia; LT=Lithuania; IT=Italy; HU=Hungary; PL=Poland; IE=Ireland; BE=Belgium; EE=Estonia; DE=Germany; FR=France; CZ=Czech Republic; SE=Sweden; UK=United Kingdom; FI=Finland; AT=Austria; NL=Netherlands

Regarding ESS, using multivariate analysis (logistic regression models with Stata software) it was possible to estimate the different attitudes among countries for an EU-wide social benefit scheme more accurately. The model has been developed for EU citizens only and includes, first of all adults' socio-demographic characteristics (age, gender, number people living in the household, citizenship, domicile), secondly, job and economic (worked or not, total household income).

In order to achieve this goal, we have used "the social benefit scheme" as the dependent variable. Social benefit scheme = 1 if the individual is in favour (otherwise against).

Concretely, in our study the following variables are considered (tab. 1):

- **Gender.** Categorical. Dummy variable: Female, Male (reference cat.).
- **Country.** Categorical. Eighteen countries. Netherlands (reference cat.), Portugal, Spain, Slovenia, Lithuania, Italy, Hungary, Poland, Ireland, Belgium, Estonia, Germany, France, Czech Republic, Sweden, United Kingdom, Finland, Austria.
- **Age group.** Categorical. Seven intervals. From 15 to 24 (reference cat.); 25 to 34; 35 to 44; 45 to 54; 55 to 64; 65 to 74; 75 +.
- **Domicile.** Categorical. Five levels. A big city (reference cat.); Suburbs or outskirts of big city; Town or small city; Country village; Farm or home in countryside.
- **Work.** Categorical. Dummy variable: Yes, No (reference cat.).

Public support for an EU-wide social benefit scheme: evidence from Round 8 of the ESS

- **Income.** Categorical. Ten levels: 1st decile (reference cat.), 2nd decile, 3rd decile, 4th decile, 5th decile, 6th decile, 7th decile, 8th decile, 9th decile, 10th decile.
- **Household.** Categorical. Five levels: 1 individual (reference cat.), 2 ind., 3 ind., 4 ind., 5 or more.

Table 1: *Logistic regression model*

SOCIAL_BENEFIT_SCHEME	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
gndr						
Female	1.096547	.0297716	3.39	0.001	1.039721	1.156479
id						
AT	.8391033	.06373	-2.31	0.021	.7230471	.9737876
BE	1.972049	.1486334	9.01	0.000	1.701228	2.285982
CZ	1.523579	.1138143	5.64	0.000	1.316068	1.763808
DE	1.532552	.1042024	6.28	0.000	1.341343	1.751019
EE	1.897868	.1387967	8.76	0.000	1.644429	2.190367
ES	6.516283	.6382684	19.14	0.000	5.378051	7.895415
FI	1.02978	.0744575	0.41	0.685	.8937143	1.18656
FR	1.429673	.1033482	4.94	0.000	1.240809	1.647284
GB	.9855533	.0736046	-0.19	0.846	.8513524	1.140909
HU	4.144868	.4159757	14.17	0.000	3.404747	5.045877
IE	2.096808	.1581692	9.82	0.000	1.80863	2.430903
IT	4.297008	.38588	16.23	0.000	3.603517	5.123961
LT	4.701796	.413748	17.59	0.000	3.956943	5.586859
PL	3.756228	.3453609	14.39	0.000	3.136819	4.497948
PT	11.19552	1.391367	19.44	0.000	8.775218	14.28337
SE	1.194457	.0921277	2.30	0.021	1.026875	1.389386
SI	6.285027	.6670041	17.32	0.000	5.104729	7.738231
cleta						
25 - 34	.8520193	.0517168	-2.64	0.008	.7564535	.9596583
35 - 44	.8226221	.0492308	-3.26	0.001	.7315757	.9249996
45 - 54	.834798	.0495328	-3.04	0.002	.7431479	.937751
55 - 64	.8623103	.0509544	-2.51	0.012	.7680077	.9681921
65 - 74	.7696442	.0480253	-4.20	0.000	.6810445	.86977
over 75	.7125496	.0489033	-4.94	0.000	.6228679	.8151439
domicil						
Suburbs or outskirts of big city	.9879447	.0518002	-0.23	0.817	.8914607	1.094871
Town or small city	.8983284	.0356845	-2.70	0.007	.8310414	.9710634
Country village	.8986947	.0363498	-2.64	0.008	.8302012	.972839
Farm or home in countryside	.7920499	.0488321	-3.78	0.000	.7018974	.8937818
pdwrk						
Marked	.8516515	.03176	-4.31	0.000	.7916236	.9162313
hinctnta						
R - 2nd decile	.9129081	.0581763	-1.43	0.153	.8057181	1.034358
C - 3rd decile	.8788089	.0561427	-2.02	0.043	.7753815	.9960324
M - 4th decile	.8422926	.0542185	-2.67	0.008	.7424565	.9555533
F - 5th decile	.745059	.0485433	-4.52	0.000	.6557402	.846544
S - 6th decile	.7814836	.0521501	-3.69	0.000	.6856735	.8906815
K - 7th decile	.7906625	.0532208	-3.49	0.000	.6929396	.902167
P - 8th decile	.6539287	.0443635	-6.26	0.000	.5725108	.7469254
D - 9th decile	.6640138	.0479113	-5.67	0.000	.5764471	.7648826
H - 10th decile	.6555659	.0483581	-5.72	0.000	.5673188	.75754
np_household						
2	1.119541	.043923	2.88	0.004	1.036681	1.209025
3	1.122155	.0546566	2.37	0.018	1.019985	1.23456
4	1.108973	.0577353	1.99	0.047	1.001396	1.228107
5 or more	1.149854	.0751645	2.14	0.033	1.011581	1.307027
_cons	1.63743	.1474701	5.48	0.000	1.372467	1.953546

Number of obs = 27,298

LR chi2(42) = 2666.19

Prob > chi2 = 0.0000
 Log likelihood = -16067.775
 Pseudo R2 = 0.0766

Source: own elaboration on QdL data (Inapp survey)

Legend: gndr=Gender; id=Country; cleta=Age Group; domicil=Domicile; pdwrk=Work;
 hinctnta=Income; np_household=Number of people living in household.

3 Conclusions

The Welfare Attitudes module shows that 67.1% of Europeans express their support for an EU-wide social benefit scheme that would guarantee a minimum standard of living for the poor.

Table 1 shows odds ratios of logistic model. The coefficients (*Beta*, not showed) can be expressed in odds by getting rid of the natural log. This is done by taking the exponential for both sides of the equation, because there is a direct relationship between the coefficients produced by logit and the odds ratios produced by logistic: a logit is defined as the natural log (base *e*) of the odds.

This fitted model says that, holding covariates at a fixed value, the odds of being in favour of a public support for an EU-wide social benefit scheme for female over the odds of being in favour of a public support for an EU-wide social benefit scheme for male (reference category) is 1.096. In terms of percent change, we can say that the odds for female are 10% higher than the odds for male. In other words, the hazard to be in favour of a public support for an EU-wide social benefit scheme is higher for female rather than male.

Regarding the citizenship, the odds of being in favour of a public support for an EU-wide social benefit scheme for all countries over the odds of being in favour of a public support for an EU-wide social benefit scheme for The Netherlands (reference category) is always higher except for Austria (OR=0.839).

The hazard to be in favour of a public support for an EU-wide social benefit scheme decreases with age (15 to 24 reference cat.), with household income (1st decile reference cat.) and for workers (not workers reference cat.).

Both attitudes are neatly aligned: in countries with strong expectations that Europeanisation will increase benefit levels, public support for an EU-level benefit scheme is comparatively strong as well.

The generosity of national welfare systems is a crucial driver of the sizeable cross-national differences in attitudes towards Social Europe. In the strongly developed Nordic welfare states, few respondents expect improvement from Europeanisation of social policy, and support for EU-level benefits is relatively low.

In the Eastern and Southern European countries, where social expenditure is considerably lower, respondents more often see the EU as an agent that could improve social protection.

Next step is to seek which variable affects the attitudes universal basic income (UBI) scheme. Using logistic regression model it is possible to estimate the different attitudes among countries more accurately.

References

1. Bonoli, G. (2010). *The Political Economy of Active Labor-Market Policy*. Politics & Society, 38(4), 435–457.
2. De Wispelaere, J., & Stirton, L. (2004). *The many faces of universal basic Income*. The Political Quarterly, 75(3), 266-274.
3. EU (2004). *The future of pension systems*. Brussels: European Commission.
4. Falkner, G. (2016). *The European Union's social dimension*. In M. Cini, & N.P.-S Borragán. (Eds.), *European Union Politics* (5th ed., pp. 275–290). Oxford: Oxford University Press.
5. Houtman, D. (1997). *Welfare State, Unemployment, and Social Justice: Judgments on the Rights and Obligations of the Unemployed*. Social Justice Research, 10 (3), 267-288.
6. Liu, X. (2016). *Applied Ordinal Logistic Regression using Stata*. Sage Publications.
7. <http://www.stata.com/bookstore/applied-ordinal-logistic-regression-using-stata>
8. Mewes, J., & Mau, S. (2013). *Globalization, socio-economic status and welfare chauvinism: European perspectives on attitudes toward the exclusion of immigrants*. International Journal of Comparative Sociology, 54(3), 228-245.
9. OECD. (2017). *Basic income as a policy option: Can it add up? Policy brief on the future of work*. Paris: OECD.

Revenue management strategies and Booking.com ghost rates: a statistical analysis

Strategie di revenue management e Booking.com ghost rates: un'analisi statistica

Cinzia Carota, Consuelo R. Nava, Marco Alderighi

Abstract To investigate hotel revenue management (RM) intensity, a dedicated database is constructed from Booking.com. A critical issue in the crawled hotel room rates is the presence of missing values for certain types of rooms, weeks of stay and booking days. Such unobserved rates are termed “ghost rates”, since they may result from RM strategies and not only from room unavailability. Our goal is to reconstruct ghost rates. To avoid bias induced by deterministic imputations, we adopt a stochastic approach to multiple imputation that exploits the time-series cross-section structure of the sampled rates and domain-specific prior knowledge, thereby improving the plausibility of imputed values and preserving, at the same time, the statistical properties of the completed data. Then, we propose a clustering of room types, based on the completed rates, useful to study RM strategies at hotel level.

Abstract *Per indagare l'intensità del revenue management (RM) nelle strutture alberghiere italiane si è costruito un database da Booking.com. Il principale problema nell'analisi delle tariffe delle camere degli hotel campionati in questo modo è la presenza di valori mancanti, per determinati tipi di camera, settimane di soggiorno e giorni di prenotazione. Tali prezzi non osservati sono detti “ghost rates”, dato che potrebbero essere il risultato di una strategia di RM e non derivare semplicemente dalla mancata disponibilità di una camera. Il nostro obiettivo è ricostruire i ghost rates. Al fine di ovviare alla distorsione indotta da tariffe imputate attraverso metodi deterministici, proponiamo un approccio probabilistico all'imputazione multipla*

Cinzia Carota

Università degli Studi di Torino, Dipartimento di Economia e Statistica, Torino (Italy) e-mail: cinzia.carota@unito.it

Consuelo R. Nava

Università della Valle d'Aosta, Dipartimento di Economia e Scienze Politiche, Aosta (Italy) e-mail: c.nava@univda.it

Marco Alderighi

Università della Valle d'Aosta, Dipartimento di Economia e Scienze Politiche, Aosta (Italy) e-mail: m.alderighi@univda.it

che sfrutta la struttura time-series cross-section dei dati e informazioni a priori specifiche di area, migliorando in tal modo la verosimiglianza delle tariffe imputate e salvaguardando, al tempo stesso, le proprietà statistiche dei dati completati. Successivamente proponiamo un clustering degli andamenti delle tariffe completate per i diversi tipi di stanza utile a studiare le strategie di RM a livello di hotel.

Key words: ghost rates; revenue management; stochastic multiple imputation; time series clustering

1 Introduction

Developed especially in airlines, hotels and rental car industries, the revenue management (RM) is a set of tools and pricing strategies designed to allocate the right capacity, to the right customer, at the right price, at the right time [7]. RM strategies are employed to take advantage from customer heterogeneity. In order to better price discriminate and, therefore, extract rent from consumers, product differentiation is often undertaken. In this frame, RM strategies also concern the use of dynamic pricing to maximize revenues [5, 10].

Only recently, hotel RM received a considerable attention, also because of the diffusion of online booking platforms such as Booking.com. Indeed, the Internet deeply changed the ways how hotels communicate and fix their room rates or room availabilities [6]. As in the case of airlines, hotel products are perishable, the room demand varies over time, and hotels have, at least in a short term, high fixed costs and low variable costs. During periods of high demand, as a results of RM strategies, rooms are usually affordable only to customers with higher willingness-to-pay, while during periods of low demand, room rates become lower.

In this paper, we aim at investigating the hotel RM activity through the study of their room rates as a function of the week of stay and the number of days between the booking and the check-in time, shortly referred to as day left. To this aim, exploiting a dedicated webcrawling system, we collected Italian hotel rates from Booking.com. Three, four and five star hotels are randomly sampled among the ones located in 22 pre-selected Italian touristic cities [4]. However, during the inspection period, missing room rates are observed. Given their numerousness, which opens challenging issues, and in order to prevent rate variability bias, we propose a method to pre-process data intended to impute missing rates.

Indeed, hotel missing rates result from two orders of reasons. On the one hand, all rooms of a given type are already occupied. On the other hand, manager decided to add or to take-off room types from the booking platform based on different motivations. First, since the commissions on Booking.com are really high, RM analysts try to maximize hotel visibility and simultaneously minimizing the sales. Second, RM analysts would prefer to sell rooms using alternative channels and only close to the date of stay they add an extra channel by putting last-minute offers on Booking.com platform. Finally, in order to offuscate the hotel room availability,

they choose to sell only a limited number of rooms on the platform and, therefore, once sold and before replaced, some room types are unavailable online. Thus, the missingness represents it-self a variety of RM strategies. Here all these missing values are termed “ghost rates” (GRs) to capture the idea that the reasons behind such unavailable room rates are unknown, but crucial to understand hotel RM tool mix and RM intensity [1].

Our major contribution is a statistical formulation of the RM analysis in hotel industry. Given that RM tools include techniques devoted to control room availability as opposed to techniques directly acting on room prices, we start our analysis by exploring the missingness map in the sample of rates crawled from Booking.com as detailed at the beginning of the next section. Preliminarily, we distinguish patterns of missingness whose generating mechanism turns out to be not at random (MNAR) from patterns whose ghost rates belong to a class of missingness ignorable from a statistical point of view.¹ For MNAR ghost rates we propose a series of ad hoc statistical models (see, e.g., [8]), also suggesting improvements in the data collection process, useful to quantify a series of specific non-pricing RM tools. For the remaining ghost rates, required for an unbiased and efficient analysis of most pricing tools, we adapt a well-established multiple imputation program (Amelia II) so as to introduce domain-specific knowledge through suitable prior distributions, in addition to smooth time trends, shifts across cross-sectional units, and correlations over time and space. Our multiple imputation model is much more flexible than the one in [1], while we consider the same ANOVA model to analyse the variability of the completed rates. Since, in general, hotel managers define specific pricing strategies for different room types, some of which with a higher dynamics in rates than others, we also suggest an appropriate clustering of hotel room types, relying on the completed data, useful to study RM at hotel level.

2 Data collection, ghost rates imputation and room types clusterization

The data collection process started the 1st of May 2018 and ended the 10th of August 2018. This webcrawling system generates a database composed of 1100 Italian hotels, for a seven-day booking period from the beginning of July 2018 to the end of October 2018. Room rates have a multiple index, h, t, d, w , with h denoting the hotel, t the room type, d the number of days left, and w the week of stay. The set of available room types in hotel h , in a given week and day left, is $C_{h,d,w} \subseteq C_h$, where C_h is the set of all room types in that hotel. $C_{h,d,w}$ is dynamically adjusted for each d and w . For each hotel, RM allocation strategies induce a three-dimensional room rate matrix, \mathbf{R} , with generic element $r_{h,t,d,w}$ and potential GRs. Indeed, our sampled \mathbf{R} exhibits 169,973 GRs (30.83%), ranging from a minimum of 0% to a maximum of

¹ Missing completely at random, MCAR, or missing at random, MAR, by applying the Little’s test and by visual inspection, respectively. The MAR assumption is also indirectly checked by using simulated missing data (see, e.g., right panels in Figure 2).

81.86% GRs per hotel. Figure 1 illustrates the great variability of room type rates in different hotels for the minimum observed day left, in the absence (hotels A and B) and in the presence (hotels C and D) of missing values with very different patterns of missingness. Ghost rates in both hotels C and D are classified as MAR, with missing depending on the week of stay (rooms booked before May 2018).

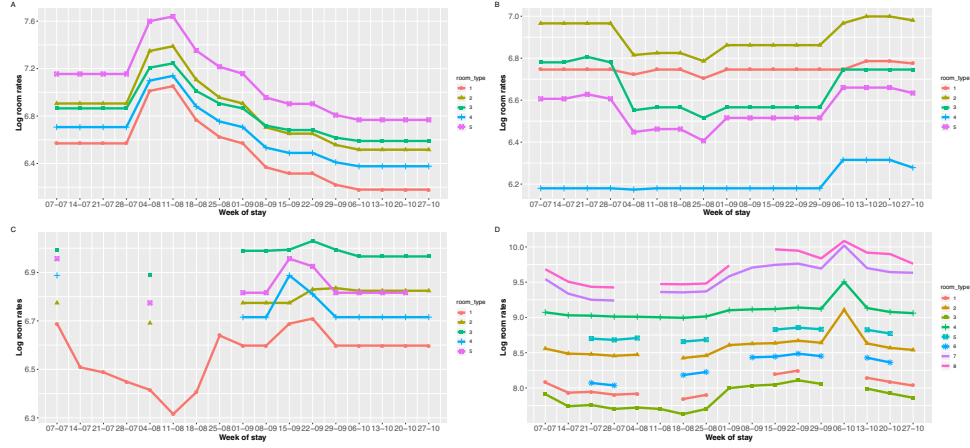


Fig. 1 Room type log-rates in four different hotels in the presence of the minimum observed day left. Hotel A is a four star hotel located in Ischia (seaside), hotel B is a four stars hotel located in Naples (art city), hotel C is a four stars hotel in Turin (art city) and hotel D is a five stars hotel located in Rome (art city).

Data imputation for multivariate time series can be a challenging problem, especially when temporal patterns as well as missingness patterns are quite different. The crude frequentist, hotel-specific approach [1] imputes GRs relying on 1100 OLS estimates:

$$\log(r_{h,t,d,w}) = \alpha_{h,1} + \beta_h x_h + \varepsilon_{h,t,d,w} \quad (1)$$

with $h = 1, \dots, H = 1100$; $t = 1, \dots, T_h$; $d = 1, \dots, D$; $w = 1, \dots, W$. The dependent variable is the natural logarithm of $r_{h,t,d,w}$, while the covariates x_h are the room type id codes represented by a set of $T_h - 1$ dummies whose effects, β_h , are additional intercepts with respect to the reference one, $\alpha_{h,1}$. Finally, $\varepsilon_{h,t,d,w}$ denotes a Gaussian white noise. Here, we enrich such imputation model in various ways, in order to take advantages of both information embedded in the entire dataset and domain-specific information. Our more flexible model,

$$\log(r_{h,t,d,w}) = \sum_{i=0}^K \beta_{h,t,i} w^i + \gamma_{h,t} d_{h,t} + \delta_{h,t} b_{h,t} + \zeta_{h,t} s_{h,t} + \eta_{h,t} f_{h,t} + L^- + L^+ + \varepsilon_{h,t,d,w}, \quad (2)$$

reduces to eq. (1) for $i = \gamma_{h,t} = \delta_{h,t} = \zeta_{h,t} = \eta_{h,t} = L^+ = L^- = 0$, $(\beta_{h,t,0} = \alpha_{h,1} + \beta_{h,t})$. For each hotel h , eq. (2) considers the multivariate, weekly-spaced, time series of log room rates, whose dimension T_h accounts for the so-called *second-degree price discrimination* while the K -th degree polynomial of the time index w is introduced to account for *peak-load pricing*. Further variables considered in the imputation model are the day left, d , to capture *inter-temporal price discrimination*, the maximum number b of guests in the room, the room size s , the free cancellation option f , together with lags and leads, L^- and L^+ respectively, of the observed log-room rate.

In addition, since hotel room minimum and maximum observed rates concur to form customer *reference prices* of a specific hotel room type, such information is suitably embedded in a prior distribution, thereby considering a Bayesian version of model (2). Finally, instability of the EM imputation algorithm is avoided by slightly shrinking the covariances among the variables toward zero by means of a so-called ridge prior.

The quality of imputations generated by our Bayesian model is explored in left panels of Figure 2, while we provide a twofold indirect check of model adequacy (and plausibility of all underlying assumptions) in right panels. There, focussing on hotel B, we show imputations of ghost rates simulated by deleting observed values so as to exactly replicate in hotel B some patterns of missingness observed in hotel D (top right panel, B_1) and by deleting completely at random 20% of the observed rates (bottom right panel, B_2). When comparing imputed rates (connected by thinner lines) with the deleted true ones (imported from Figure 1, panel B), despite the low degree of the polynomial function ($K=1$), we observe quite good imputation results in both B_1 and B_2 cases.

We then apply to the completed time series of rates a clustering of room types governed by a dissimilarity measure able to capture and compare the higher-level dynamic structures describing the global behaviour of the series. In particular, to group homogeneous rate patterns useful to study RM activity at hotel level, we realize a hierarchical clustering based on the Ward's method and a distance constructed by considering the partial autocorrelation functions with geometric weights decaying with the lag [9]. The selection of this distance ensures the reasonable partition of hotel rooms presented in Figure 3. The goal of this clustering is to group rooms with affine rate patterns, not necessarily associated with similar rate levels. Indeed, different and sometime counterintuitive clusters are obtained when considering distances based on raw data, on correlation or on discrete wavelet transform.

References

1. Alderighi, M., Calabrese, M., Christille, J.M., Nava, C.R., Salvemini, C.: Room Rates and Hotel Price Fairness, mimeo (2019)
2. Bauer, J., Angelini, O., Denev, A.: Imputation of Multivariate Time Series Data - Performance Benchmarks for Multiple Imputation and Spectral Techniques (July 3, 2017). Available at SSRN: <https://ssrn.com/abstract=2996611> or <http://dx.doi.org/10.2139/ssrn.2996611>

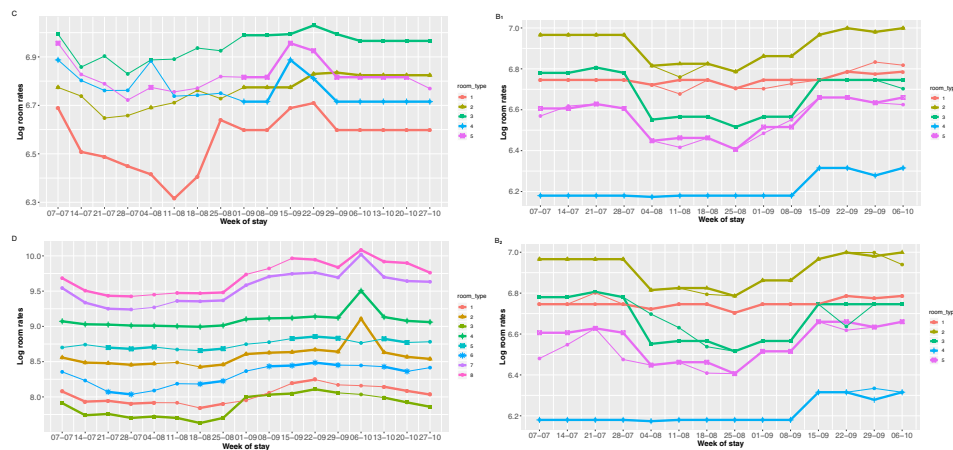


Fig. 2 GR imputations according to the Bayesian version of model (2) in hotels C and D (left panels) and in hotel B (right panels), in the presence of GRs generated so as to replicate in that hotel five patterns of missingness observed in hotel D (panel B_1) or completely at random (panel B_2). In all panels, imputed rates are connected to the observed ones by thinner lines.

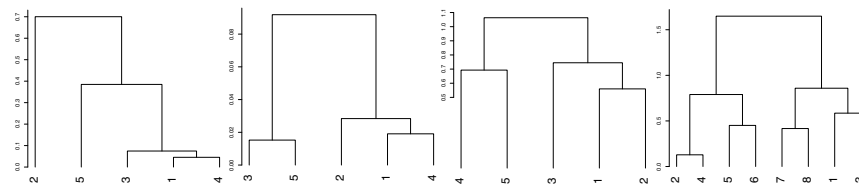


Fig. 3 Dendrogram obtained from a hierarchical clustering of completed room rates multiple imputation in hotels A, B, C and D.

3. Honaker, J., King, G., Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of statistical software*, 45(7), 1-47
4. ISTAT: Movimento turistico in Italia. ISTAT, Statistiche - report, 1-22 (2017)
5. Ivanov, S., Zhechev, V.: *Hotel revenue management: from theory to practice*. Varna: Zangador (2014)
6. Kannan, P. K. K.: Dynamic pricing on the Internet: Importance and implications for consumer behavior. *International Journal of Electronic Commerce*, 5(3), 63-83 (2001)
7. Kimes, S. E. and Wirtz, J.: Has revenue management become acceptable? Findings from an international study on the perceived fairness of rate fences. *Journal of Service Research*, 6(2), 125-135 (2003)
8. Molenberghs, G., Verbeke, G.: *Models for discrete longitudinal data*, New York: Springer-Verlag (2005)
9. Montero, P., Vilar, J. A.: TSclust: An R package for time series clustering. *Journal of Statistical Software*, 62(1), 1-43 (2014)
10. Talluri, K. T., van Ryzin, G. J.: *The theory and practice of revenue management*. International series in operations research & management science. Kluwer Academic Publishers, Boston, MA, 2-14 (2004)

Analysing international migration flows: a Bayesian network approach

Analisi dei flussi migratori internazionali attraverso l'impiego di modelli grafici

Federico Castelletti and Emanuela Furfaro

Abstract In this paper we investigate the non-economic determinants of international migration flows. We approach this problem using a Bayesian graphical model methodology which performs variable selection among a set of covariates that are potentially related to migration flows. We consider inflows to Italy from 171 different countries of origin on which we measure demographic and geographic characteristics. While results are coherent with the most recent literature, our method also provides measures of uncertainty around the dependence structure between covariates and migration flows.

Abstract Nel presente lavoro si analizzano le determinanti di carattere “non economico” che influenzano le migrazioni internazionali. Il problema è affrontato utilizzando una metodologia bayesiana basata su modelli grafici che seleziona variabili rilevanti a partire da un insieme di covariate tipicamente collegate ai flussi migratori. Si considerano i flussi in ingresso verso l'Italia da 171 Paesi di origine sui quali sono rilevate caratteristiche demografiche e geografiche. I risultati, coerenti con la letteratura recente, forniscono inoltre una misura di incertezza sulla struttura di dipendenza tra covariate e flussi migratori.

Key words: migration flows, model selection, Bayesian networks.

Federico Castelletti
Dipartimento di Scienze Statistiche, Largo Gemelli 1, Milano
e-mail: federico.castelletti@unicatt.it

Emanuela Furfaro
Dipartimento di Scienze Statistiche, Largo Gemelli 1, Milano
e-mail: emanuela.furfaro@unicatt.it

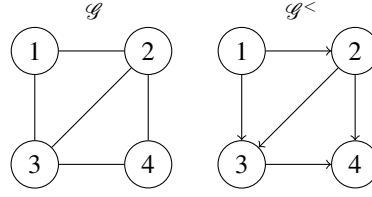
1 Introduction

In the last decades, international migrations towards developed countries have significantly grown thus making prediction of migratory flows a non-negligible component in population size projections [1]. Determinants of international migrations firstly include economic and political issues. Coherently, most of the recent literature has focused on the effects of economic and political variables [4], while only few studies have also accounted for non-economic (e.g. demographic) factors [5]. However, demographic variables are quite easy to predict and thus an extended analysis that also incorporates such characteristics can significantly improve the prediction of international migrations [5].

In this paper we investigate the relationship between non-economic variables and international migrations using a graphical model based approach. Graphical models represent a promising and effective tool for discovering dependence relationships among potentially many variables. From a statistical point of view, we consider a problem of covariate selection for a response variable inferring a graphical structure which jointly models the dependence relationships within covariates and between covariates and response.

2 Methods

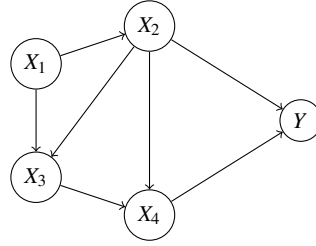
We first introduce some general notation. A graph \mathcal{G} is a pair (V, E) where $V = \{1, \dots, p\}$ is a set of vertices (or nodes) and $E \subseteq V \times V$ a set of edges. Let $u, v \in V$, $u \neq v$. If $(u, v) \in E$ and $(v, u) \notin E$ we say that \mathcal{G} contains the directed edge $u \rightarrow v$. If instead $(u, v) \in E$ and $(v, u) \in E$ we say that \mathcal{G} contains the undirected edge $u - v$. Two vertices u, v are adjacent if they are connected by an edge (directed or undirected). For any pair of distinct nodes $u, v \in V$, we say that u is a *parent* of v if $u \rightarrow v$. Conversely, we say that v is a *son* of u . The set of all parents of u in \mathcal{G} is denoted by $\text{pa}_{\mathcal{G}}(u)$. A graph is called *directed* (*undirected*) if it contains only directed (undirected) edges. A directed graph is called Directed Acyclic Graph (DAG for short, denoted by \mathcal{D}) if it does not contains cycles, that is a sequence of edges $v_1 \rightarrow v_2 \rightarrow \dots \rightarrow v_k$ such that $v_1 \equiv v_k$. A particular class of undirected graphs is represented by *decomposable* graphs, also called *chordal* or *triangulated*. An undirected graph is decomposable if every path of length $l \geq 4$ contains a chord, that is two non-consecutive adjacent vertices [6]; see for instance graph \mathcal{G} in Figure 1. A graph encode a set of (marginal and) conditional independencies which determines its Markov property and can be read off from the graph itself using the notion of *d-separation* [7]. Moreover, we say that two graphs are *Markov equivalent* if and only if they encode the same conditional independencies. Markov equivalent graphs are not distinguishable in the presence of observational data only (in other terms they are “score equivalent”). Most importantly, for each decomposable undirected graph \mathcal{G} we can find a perfect directed version, $\mathcal{G}^<$ (a DAG), which is Markov equivalent to \mathcal{G} [6]; see also Figure 1.

Fig. 1: A decomposable undirected graph \mathcal{G} and its perfect directed version $\mathcal{G}^<$.

Let now Y be a response variable, X_1, \dots, X_q a collection of covariates. Each variable (both the response and covariates) can be associated to a node in a graph, whose structure will constrain the sampling distribution of the data. We are interested in selecting which covariates *directly* affect the response. In addition we allow for the presence of a dependence structure among covariates that we model by means of an undirected decomposable graph $\mathcal{G}_x = (V_x, E_x)$ (within covariates graphical structure) where $V_x = \{x_1, \dots, x_q\}$ and $E_x \subseteq V_x \times V_x$ is the set of (undirected) edges between covariates. We also denote with $\mathcal{G}_x^<$ the perfect directed version of \mathcal{G}_x . Dependence relationships between covariates and response are instead represented by a directed graph $\mathcal{D}_{y|x} = (V_{xy}, E_{y|x})$ where $V_{xy} = \{x_1, \dots, x_q, y\}$ and $E_{y|x} \subseteq \{x_1, \dots, x_q\} \times y$. Consequently, in $\mathcal{D}_{y|x}$ we allow for the presence of directed edges *from* the covariates *to* the response only. The entire graphical structure, that we call *regression DAG*, is finally determined by the union of $\mathcal{G}_x^<$ and $\mathcal{D}_{y|x}$ and is denoted by \mathcal{D}_{xy} (or simply \mathcal{D} as in the sequel); see for instance Figure 2. For a given DAG \mathcal{D} we can write the factorization

$$f(x_1, \dots, x_q, y) = \prod_{j=1}^q f(x_j | \mathbf{x}_{\text{pa}_{\mathcal{D}}(j)}) \cdot f(y | \mathbf{x}_{\text{pa}_{\mathcal{D}}(y)}), \quad (1)$$

where $\text{pa}_{\mathcal{D}}(j)$ is the set of parents of node j in \mathcal{D} .

Fig. 2: A regression DAG with $q = 4$ covariates X_1, \dots, X_4 ; the within covariates graphical structure corresponds to the perfect directed version of the decomposable graph \mathcal{G} in Figure 1.

In a Gaussian framework we now assume $X_1, \dots, X_q, Y | \Omega_{\mathcal{D}} \sim \mathcal{N}(\mathbf{0}, \Omega_{\mathcal{D}}^{-1})$, where $\Omega_{\mathcal{D}}^{-1}$ is the precision matrix (inverse of the covariance matrix $\Sigma_{\mathcal{D}}$) Markov w.r.t. DAG \mathcal{D} and hence *constrained* by \mathcal{D} . Let also \mathbf{y} be a $(n, 1)$ vector of observations from the response Y , \mathbf{X} a (n, q) data matrix collecting the observations from the q covariates. We denote with \mathcal{S} the set of all regression DAGs on $q + 1$ nodes which will represent our model space. The objective is then to perform model selection within the space of regression DAG models given the observed data (\mathbf{X}, \mathbf{y}) ; we approach such problem adopting a Bayesian methodology. Specifically, let $f(\mathbf{y}, \mathbf{X} | \Omega_{\mathcal{D}})$ be the likelihood function, $p(\Omega_{\mathcal{D}})$ a prior assigned to the DAG model parameter $\Omega_{\mathcal{D}}$. We are interested in evaluating the marginal likelihood $m(\mathbf{y}, \mathbf{X} | \mathcal{D})$ of a generic $\mathcal{D} \in \mathcal{S}$ which from a Bayesian perspective represents the *score* assigned to model \mathcal{D} ,

$$m(\mathbf{y}, \mathbf{X} | \mathcal{D}) = \int f(\mathbf{y}, \mathbf{X} | \Omega_{\mathcal{D}}) p(\Omega_{\mathcal{D}}) d\Omega_{\mathcal{D}}. \quad (2)$$

To this end we rely on the *objective Bayes* method of [3] who derive a closed formula for $m(\mathbf{y}, \mathbf{X} | \mathcal{D})$. Let now $p(\mathcal{D})$ be a prior assigned to \mathcal{D} . Bayesian prior-to-posterior analysis amounts to evaluate the posterior probability of \mathcal{D} given the data,

$$p(\mathcal{D} | \mathbf{y}, \mathbf{X}) = \frac{m(\mathbf{y}, \mathbf{X} | \mathcal{D}) p(\mathcal{D})}{\sum_{\mathcal{D} \in \mathcal{S}} m(\mathbf{y}, \mathbf{X} | \mathcal{D}) p(\mathcal{D})}, \quad (3)$$

for each $\mathcal{D} \in \mathcal{S}$; see also [2]. Since an exhaustive enumeration of *all* the regression DAGs on $q + 1$ nodes is not feasible, we construct a Markov chain Monte Carlo (MCMC) algorithm to traverse the model space and approximate the posterior distribution in Eq. 3. Our MCMC method is based on a Markov chain on the model space which performs moves between graphs through additions and removals of edges provided that each proposed graph *falls inside* the model space (equivalently it must be a regression DAG); see also [2] for a general theoretical framework. The output of our MCMC consists in a collection of regression DAGs visited by the Markov chain at each time. Accordingly, the posterior distribution in (3) is approximated by the number of visits of each model. In addition we can compute the posterior probability of inclusion of each edge and obtain a single model estimate, if required, by selecting those edges whose posterior probability is greater than some threshold (e.g. 0.5); see also [2] for details and the output of Figure 3.

3 Application

We consider Italy as destination country and inflows from 171 different origin countries at 3 different time spans, i.e. 2000, 2010 and 2016. In this first analysis, we consider each year separately. The response variable Y is then the logarithm of the

annual number of migrants from origin country i to Italy in a given year t ¹. The set of covariates includes the following characteristics of the origin country:

- X_1 : total population,
- X_2 : percentage of urban population,
- X_3 : Potential Support Ratio (PSR), defined as the ratio of people younger than 15 to the working-age population (those aged 15-64),
- X_4 : Infant Mortality Rate (IMR), defined as the probability of a live birth to die before one year of age,
- X_5 : distance between the capital of origin country and Italy.

The dataset contains no missing data and all variables were zero-centred. The normality assumption is reasonably satisfied after log-transformations.

For the sake of brevity, only results for the latest year considered (2016) are presented. Similar results, available upon request from the Authors, were obtained for years 2000 and 2010. Results are summarized in Figure 3. The left panel contains the (5,5) heatmap with marginal posterior probabilities of edge inclusion for the decomposable within covariates graphical structure and the (5,1) heatmap with probabilities of inclusion for the directed edges between covariates and response. In the right panel we instead report the median probability graph model, which is obtained by selecting those edges whose posterior probability of inclusion is greater than 0.5. With regards to the within covariates structure it appears that PSR (X_3) is clearly related to the percentage of urban population (X_2) and IMR (X_4). Such result is very reasonable as all these variables concern development and economic conditions. Moreover, among the covariates, only X_1 , X_3 and X_4 *directly* affect the response. Consequently, we would say that the effect of the other covariates on the migration flow is “filtered” by them. For instance, using a more technical terminology, Y is conditionally independent of X_2 given $\{X_3, X_4\}$, $Y \perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$.

4 Discussion

Graphical models represent an effective and powerful tool to study dependencies between variables and provide results that are easy and straightforward to interpret. In addition our methodology, being fully Bayesian, returns a posterior distribution over the space of *all* possible regression DAG models. This in turn provides a coherent quantification of any measure of uncertainty around the *strength* of the dependence relationship between variables. For simplicity we considered in our study only few independent variables, but many other demographic characteristics can be included. Our results are coherent with the literature; in addition we explored the underlining structure between the determinants of international migrations.

¹ Data Sources: Y , OECD Stat, <https://stats.oecd.org>; $X_1 - X_4$, The Worldbank Database, <https://data.worldbank.org/>; X_5 , Centre d'Etudes Prospectives et d'Informations Internationales (CEPII), <http://www.cepii.fr/>

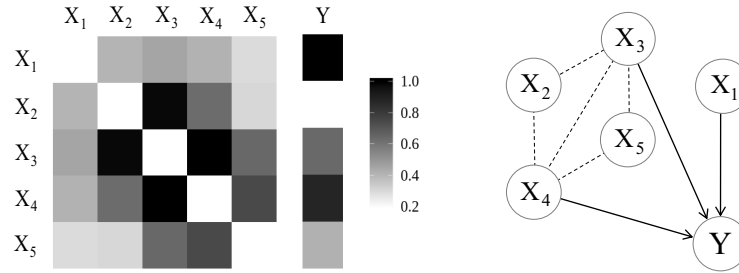


Fig. 3: Heatmaps with marginal posterior probabilities of edge inclusion for the graphical structure within covariates and between covariates and response (left panel). Median probability graph model (right panel).

Alternative techniques to investigate dependency relationships between variables are of course present in the literature. Among these, Structural Equation Modelling (SEM) and path analysis are the most used. However, while path analysis typically requires causal assumptions underlying the dependencies between variables and aim at estimating the size of such causal relationships under a given path diagram, our approach is more targeted to discover conditional independencies between variables and hence more general. Moreover, differently from SEM, the proposed method implicitly assumes that there are no latent variables in the system.

This contribution clearly presents some limitations which give room for future improvements. First, we analysed the data for each year separately without accounting for the effect of time over migration flows. Furthermore, we based our study on a Bayesian methodology for model selection of Gaussian graphical models which cannot be easily adapted to include different types of variables (e.g. categorical), such as the presence of colonial links or a common language between countries which have also proven to be important determinants of international migrations.

References

1. Azose, J.J., Raftery, A.E.: Estimating large correlation matrices for international migration. *The Annals of Applied Statistics* **12** (2), 940–970 (2018)
2. Castelletti, F.: Bayesian model selection of Gaussian DAG models. Under review (2019)
3. Consonni, G., La Rocca, L.: Objective Bayes Factors for Gaussian Directed Acyclic Graphical Models. *Scandinavian Journal of Statistics*. **39** (4), 321–354 (2012)
4. Fertig, M., Schmidt, C. M.: Aggregate-Level Migration Studies as a Tool for Forecasting Future Migration Streams. *IZA Discussion Paper* **183** (2000)
5. Kim, K., Cohen, J.E.: Determinants of International Migration Flows to and from Industrialized Countries: A Panel Data Approach Beyond Gravity. *International Migration Review*. **44** (4), 899–932 (2010)
6. Lauritzen, S. L.: *Graphical Models*. Oxford University Press, Oxford (1996)
7. Pearl, J.: *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge (2000)

A sparse estimator for the function-on-function linear regression model

Uno stimatore sparso per il modello di regressione lineare con regressore e risposta funzionali

Fabio Centofanti, Matteo Fontana, Antonio Lepore, and Simone Vantini

Abstract A new estimator for the non-concurrent function-on-function linear regression model is proposed that is able to better identify the part of the domain where the coefficient function is exactly zero and thus to increase the interpretability of the model. This is achieved by introducing a functional LASSO penalization and two roughness penalties that ensure sparseness and smoothness of the estimate, respectively. By means of a wide Monte Carlo simulation study, the interpretability as well as estimation and predictive performances of the proposed estimator are eventually discussed.

Abstract *In questo lavoro viene proposto un nuovo stimatore per il modello di regressione lineare con regressore e risposta funzionali, in grado di migliorare l'identificazione delle parti di dominio dove il coefficiente di regressione è esattamente zero. Tale proprietà (sparseness), che garantisce una maggiore interpretabilità del modello, viene ottenuta mediante un opportuno termine di penalizzazione LASSO funzionale. In aggiunta, al fine di conferire liscezza (smoothness) allo stimatore proposto, vengono introdotti ulteriori due termini di penalizzazione della rugosità. Mediante simulazione Monte Carlo vengono infine discusse le proprietà di stima e previsione dello stimatore proposto.*

Key words: functional data analysis, linear regression, LASSO, B-splines

Fabio Centofanti, Antonio Lepore

Dept. of Industrial Engineering, University of Naples Federico II, P.le V. Tecchio 80, 80125, Naples, Italy

e-mail: fabio.centofanti@unina.it, antonio.lepore@unina.it

Matteo Fontana

Dept. of Management, Economics, and Industrial Engineering, Politecnico di Milano, Italy.

e-mail: matteo.fontana@polimi.it

Simone Vantini

MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy

e-mail: simone.vantini@polimi.it

1 Methods

Functional linear regression (FLR) is the generalization of the classical regression analysis to the context of the functional data analysis (FDA) [7, 2, 3, 5]. In particular, the non-concurrent function-on-function linear regression model is studied in this paper, that is a FLR model where both the regressor and response variable have a functional form. Let the pairs (X_i, Y_i) , $i = 1, \dots, n$, be independent realizations of regressor X and response Y , which are assumed to be smooth random processes with mean zero defined on the compact domains \mathcal{T} and \mathcal{S} . Then, the non-concurrent function-on-function linear regression model can be expressed in the following form

$$Y_i(t) = \int_{\mathcal{S}} X_i(s) \beta(s, t) ds + \varepsilon_i(t) \quad t \in \mathcal{T}, \quad (1)$$

where ε_i are independent random functional errors with mean zero and variance function σ^2 and are independent of X_i , $i = 1, \dots, n$. The bivariate function β is defined on $\mathcal{T} \times \mathcal{S}$, and hereinafter referred to as *coefficient function*.

In this work, we analyse the estimator named S-LASSO (Smooth plus LASSO) of the coefficient function β proposed in [1] that is able (i) to better identify the region where β is zero (*null region*) and (ii) to estimate β on the region where it is different from zero (*non-null region*). The S-LASSO estimator $\hat{\beta}_{SL}$ is defined as the solution of an optimization problem whose objective function, to be minimized, is composed by the usual sum of squared errors added to two roughness penalties, which are applied to β only as a function of s and t , respectively, and a functional LASSO penalty. To compute $\hat{\beta}_{SL}$, we consider the space generated by two sets of B-spline basis functions. With this restriction, the problem reduces to a multivariate optimization problem and, thus, standard optimization algorithms can be used. – In [1] the S-LASSO estimator is proved to have good theoretical properties. In particular, under some regularity conditions, it is consistent (i.e., $\hat{\beta}_{SL}$ converges to β in probability) and pointwise sign consistent (i.e., $\text{sign}(\hat{\beta}_{SL}(s, t))$ converges to $\text{sign}(\beta(s, t))$ in probability $\forall (s, t) \in [\mathcal{S} \times \mathcal{T}]$).

2 Simulation Study

We conduct a Monte Carlo simulation study to explore the performance of the proposed estimator in the two following scenarios, defined by the form of the coefficient function β .

Scenario I β is zero over the whole domain.

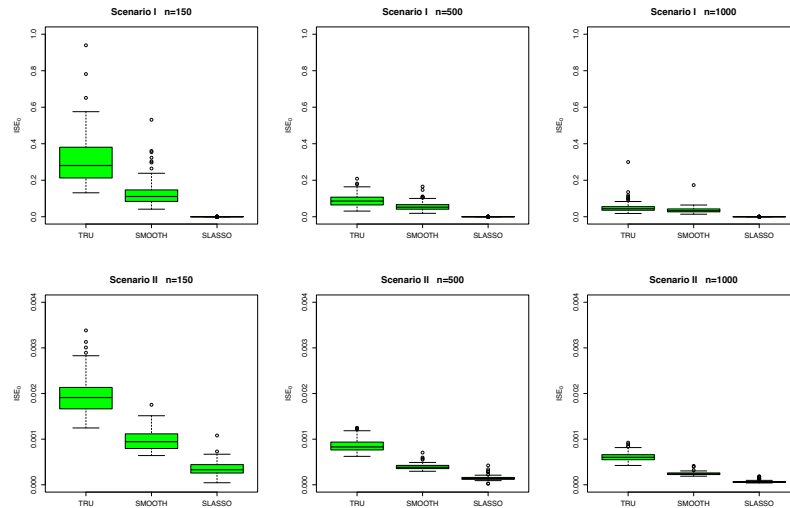
Scenario II β is different from zero in the central part of the domain.

For each Scenario, and for $n = 150, 500, 1000$, we generate 100 datasets composed by a training set of n observations to estimate $\hat{\beta}_{SL}$ and a test set with size equal to 4000 to test its predictive performance.

As in [6], we consider the integrated squared error (ISE) to assess the quality of the estimator. In particular, we denote with ISE_0 and ISE_1 the ISE over the null and the non-null regions, respectively. Moreover, we consider the prediction mean squared error (PMSE) to quantify predictive performances. By means of those indices, the S-LASSO estimator is compared with the two estimators proposed by [7], which are both assumed in a finite dimension tensor space and with regularization achieved either by choosing the dimension of the tensor space or by introducing roughness penalties. They will be referred to as TRU (truncation) and SMOOTH methods, respectively. TRU, SMOOTH and S-LASSO estimators are all computed by using cubic B-spline basis functions with evenly spaced knot sequence.

From Figure 1 we observe that the estimation error on the null region (in terms of ISE_0) of the S-LASSO estimator is significantly smaller than those of the other estimators in both scenarios. In particular, for Scenario I, the ISE_0 of the proposed estimator is practically zero, also when $n = 150$. Whereas, TRU and SMOOTH estimators are more sensitive to the sample size.

Fig. 1 The integrated squared error on the null region (ISE_0) for the TRU, SMOOTH and S-LASSO estimators in Scenario I and II.



Conversely, Figure 2 shows the S-LASSO estimator to be outperformed on the non-null region by the SMOOTH estimator for all sample sizes. As the multivariate LASSO method is known to overshrink the estimator of the non-null coefficients [4], we surmise that this phenomenon arises in the function-on-function linear model as well. Lastly, Figure 3 displays the S-LASSO estimator to achieve the lower PMSE in both scenarios, especially at $n = 150$. Differences with competitor estimators tend to be narrower as the sample size increases.

Fig. 2 The integrated squared error on the non-null region (ISE_1) for the TRU, SMOOTH, and S-LASSO estimators in Scenario II.

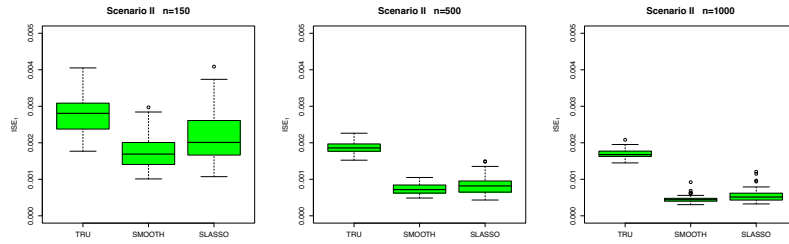
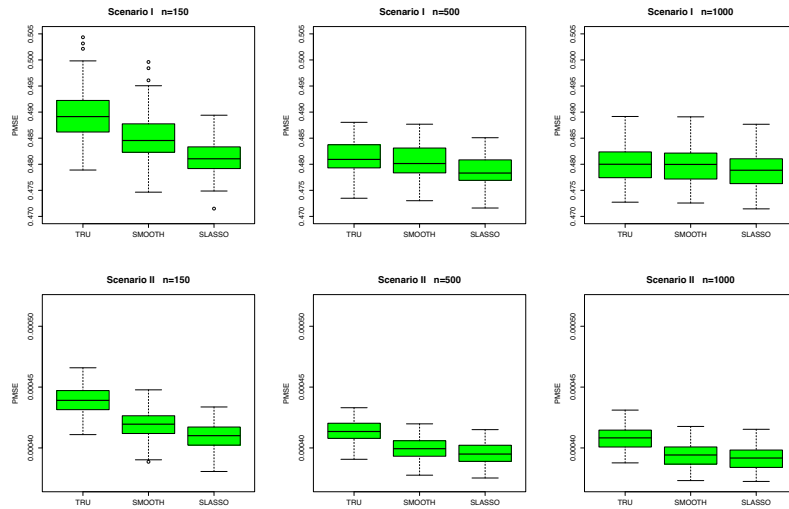


Fig. 3 The prediction mean squared error (PMSE) for the TRU, SMOOTH and S-LASSO estimators in Scenario I and II.



3 Conclusion

The LASSO is one of the most used and popular method to estimate the linear regression coefficients, because it is able to ensure prediction accuracy while performing variable selection. The proposed S-LASSO estimator is the combination of three parts: the functional LASSO penalty, which has the task of shrinking toward zero the coefficient function estimator on the null region; the B-spline basis functions and the roughness penalties, which are needed to make the estimator smooth. By comparison with other estimators proposed in the literature, the S-LASSO is shown, through a wide Monte Carlo simulation, to improve the model interpretability and to guarantee good estimation and predictive performances.

References

1. Centofanti, F., Fontana, M., Lepore, A., Vantini, S.: Smooth lasso estimator for function-on-function linear regression model. Manuscript - MOX report (2019)
2. Horváth, L., Kokoszka, P.: Inference for functional data with applications, vol. 200. Springer Science & Business Media (2012)
3. Hsing, T., Eubank, R.: Theoretical foundations of functional data analysis, with an introduction to linear operators. John Wiley & Sons (2015)
4. James, G.M., Radchenko, P.: A generalized dantzig selector with shrinkage tuning. *Biometrika* **96**(2), 323–337 (2009)
5. Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. CRC Press (2017)
6. Lin, Z., Cao, J., Wang, L., Wang, H.: Locally sparse estimator for functional linear regression models. *Journal of Computational and Graphical Statistics* **26**(2), 306–318 (2017)
7. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer Series in Statistics. Springer (2005)

Robustness and fuzzy multidimensional poverty indicators: a simulation study

Robustezza ed indicatori fuzzy multidimensionali della povertà: uno studio di simulazione

Michele Costa

Abstract This paper proposes a simulation study in order to evaluate the robustness of fuzzy sets indicators applied to the poverty measurement. We address the issues related to the subjectivity which affects the choice of membership to the poor set. The subjective choices of the individual researchers could lead to unstable results and then to a lack of robustness of the method. We investigate the effects of the subjectivity by means of a Monte Carlo study and we provide evidence of an extremely satisfactory robustness level for fuzzy multidimensional poverty indicators.

Abstract *In questo lavoro viene proposto uno studio di simulazione per valutare la robustezza degli indicatori fuzzy applicati alla misura della povertà. Si analizza, in particolare, la soggettività legata alla scelta dei livelli di appartenenza all'insieme dei poveri. Le scelte soggettive dei singoli ricercatori potrebbero portare a risultati instabili e quindi ad una mancanza di robustezza. Gli effetti della soggettività vengono analizzati grazie ad uno studio Monte Carlo che suggerisce complessivamente un livello di robustezza estremamente soddisfacente.*

Key words: Fuzzy indicators, Multidimensional poverty index, Simulation study

1 Introduction

Poverty is a multidimensional condition, not exclusively related to an insufficient provision of income or wealth, but the result of the contemporaneous occurrence of many factors [5], [7]. A relevant implication of the multidimensional view is the importance to divide total population not simply into just two subgroups, poor and non poor, as it is done with poverty lines, but to detect a multiplicity of populations subgroups, each related to different poverty levels, gradually moving from the completely poor to the absolutely non poor condition. Only 2 subgroups represent a

Michele Costa

Department of Economics, University of Bologna, e-mail: michele.costa@unibo.it

severe limit with just one inequality factor, but they would be absolutely inadequate by using many inequality factors.

A simple but effective method to detect different degrees of poverty, overcoming the dichotomy poor and non poor, is given by the fuzzy sets. In this work we aim to address some issues related to the choice of the degrees of membership to the poor of the different population subgroups: this choice belongs to the researchers and, therefore, is frequently view as arbitrary and questionable. The weakness implied by subjective choices refers to their influence on the stability of the results, that is, on the robustness of the method. In the following we propose a simulation study in order to assess the effects on the poverty indicator of different degrees of membership and to evaluate the robustness of the fuzzy poverty indicators.

2 A fuzzy multidimensional poverty indicator

Many Authors proposed, generalized or applied fuzzy sets based methods to the poverty analysis, among the others [3], [1], [4], [6]. Let be A the set of the n units, B the subset of the poor units, and $\mu(X)$ an indicator function, which determines the degree of membership of the i -th unit to the set B .

Usually $\mu(X)$ can assume only two values, 1 for the poor units and 0 for the non poor units, but, within the fuzzy sets, $\mu(X)$ ranges between 0 and 1, thus allowing intermediate or partial membership to B : if $\mu(X)_i = 1$ and $\mu(X)_i = 0$ still indicate a poor and a non poor unit, respectively, values as $\mu(X)_i = 0.8$ and $\mu(X)_i = 0.2$ refer to an almost poor and to an almost non poor unit, and $\mu(X)_i = 0.5$ to a neither poor or non poor unit. A finer classification with respect to only two cases, 0 and 1, is a relevant improvement, but the choice of the intermediate values depends on the individual researchers, thus introducing into the analysis a certain degree of subjectivity and, potentially, of uncertainty and unstability.

The presence of m inequality factors leads to introduce a second subscript j , where $\mu(X_j)_i$ denotes the degree of membership to B of the i -th unit with respect to the j -th inequality factor. Given n units and m inequality factors, we obtain a matrix $M(X)_{n \times m}$ of $\mu(X_j)_i$. By columns of the matrix $M(X)$ we derive an unidimensional index, which evaluates the poverty condition with respect to a certain attribute X_j :

$$I(X_j) = \sum_{i=1}^n \mu(X_j)_i n_i / \sum_{i=1}^n n_i. \quad (1)$$

The traditional head count ratio is a special case of $I(X_j)$, obtained when the indicator function can be only 0 or 1 and the population is divided in two subgroups, poor and non poor, on the basis of a poverty line z :

$$\mu(X) = \begin{cases} 1 & \text{if } x_i < z \\ 0 & \text{if } x_i \geq z. \end{cases} \quad (2)$$

In the framework of fuzzy sets, the indicator function can assume multiple values, $\mu(X) = \{\mu_1, \mu_2, \dots, \mu_k\}$, with $\mu_1 = 1$, $\mu_k = 0$ and the intermediate values chosen by the researchers. By rows of the matrix $M(X)$ we obtain a multidimensional index for the i -th unit, calculated as a weighted sum of the $\mu(X_j)_i$:

$$I(a_i) = \sum_{j=1}^m \mu(X_j)_i w_j / \sum_{j=1}^m w_j.$$

The weights w_j measure the intensity of deprivation and social exclusion related to the j -th inequality factor: the underlying guideline is that a factor not possessed by any unit has no effect on the social exclusion, while, on the contrary, if all the units but a few possess the factor, it represents a relevant source of inequality and social exclusion. A weighting system consistent with that is [2]

$$w_j = \log(n / \sum_{i=1}^n \mu(X_j)_i n_i)$$

where w_j is equal to 0 for $\mu(X_j)_i = 1$, $i = 1, \dots, n$, that is when the j -th factor is not possessed by any unit, while w_j increases for decreasing $\sum_{i=1}^n \mu(X_j)_i$.

It is relevant to observe how the unidimensional framework, or the exclusion of a poverty attribute X_j , corresponds to an indicator function $\mu(X_j) = 1$, which implies to assume that all units have the same value of X_j .

The synthesis of the m unidimensional indexes $I(X_j)$ and of the n multidimensional indexes $I(a_i)$ in a multidimensional poverty index can be achieved quite simply by means of a weighted sum

$$I = \frac{\sum_{j=1}^m I(X_j) w_j}{\sum_{j=1}^m w_j} = \frac{\sum_{i=1}^n I(a_i) n_i}{\sum_{i=1}^n n_i} = \frac{\sum_{j=1}^m w_j \sum_{i=1}^n \mu(X_j)_i n_i}{\sum_{j=1}^m w_j \sum_{i=1}^n n_i} \quad (3)$$

One of the issues about fuzzy sets based indicators, which is also frequently perceived as one of its weaknesses, refers to the subjectivity of the values attributed to the indicator function $\mu(X)$. In order to investigate this point in the following we develop a simulation study.

3 The simulation study for fuzzy indicators

With the aim to assess the degree of robustness of the method and to evaluate the effects of $\mu(X)$ on the poverty index, we propose a simulation study where the scores of the indicator function are not a priori chosen, but endogenously determined by means of a Monte Carlo experiment. We randomly extract $\mu(X)$ from an uniform distribution within an interval determined according to the following scheme:

$$\mu(X_j) = \begin{cases} \mu_{j1} & \text{with } \mu_{j1} \in [\mu_{j2} + h, 1] \\ \dots & \\ \mu_{jk-1} & \text{with } \mu_{jk-1} \in [\mu_{jk} + h, 2/k] \\ \mu_{jk} & \text{with } \mu_{jk} \in [0, 1/k] \end{cases} \quad (4)$$

where k is the exogenous number of classes and h is an exogenous minimum distance between two different values of $\mu(X)$.

As a case study for our simulation study we resort to the income distribution of the Italian households for the 2016. We illustrate the advantages of fuzzy indicators starting from the unidimensional case by comparing the traditional head count ratio to a fuzzy indicator and to a simulated case. First we analyse an income based poverty indicator (X_1 = household equivalent income): the headcount ratio, obtained as (1), with $\mu(X_1)$ equal to (2) and z being the 60% of the median \bar{x}_{me} of the equivalent income: $z = 0.6 * \bar{x}_{me}$.

An income based fuzzy poverty indicator, instead to divide total population in only two subgroups, is based on $k > 2$ subgroups: for example

$$\mu(X_1) = \begin{cases} 1.0 & \text{if } x_{1i} < 0.4\bar{x}_{1me} \\ 0.9 & \text{if } 0.4\bar{x}_{1me} \leq x_{1i} < 0.6\bar{x}_{1me} \\ 0.5 & \text{if } 0.6\bar{x}_{1me} \leq x_{1i} < 0.8\bar{x}_{1me} \\ 0 & \text{if } 0.8\bar{x}_{1me} \leq x_{1i} \end{cases} \quad (5)$$

In (5) we allow a greater flexibility with respect to (2): under the poverty line z we assume two classes, introducing a difference among the poor, and also above the poverty line we differentiate among the non poor, assigning a positive membership ($\mu_{13} = 0.5$) to the units with income lesser than $0.8\bar{x}_{1me}$.

In order to vary the scores of the indicator function $\mu(X_1)$, we run a Monte Carlo experiment. By setting $k = 4$ and $h = 0.1$, on the basis of (4) we obtain

$$\mu(X_1) = \begin{cases} \mu_{11} & \text{with } \mu_{11} \in [\mu_{12} + 0.1, 1] \\ \mu_{12} & \text{with } \mu_{12} \in [\mu_{13} + 0.1, 0.75] \\ \mu_{13} & \text{with } \mu_{13} \in [\mu_{14} + 0.1, 0.50] \\ \mu_{14} & \text{with } \mu_{14} \in [0, 0.25] \end{cases} \quad (6)$$

that is we extract μ_{14} , the degree of membership to the poorest subgroup, from an uniform distribution within the interval $[0, 0.25]$, then we move to μ_{13} , which belongs to the interval $[\mu_{14} + 0.1, 0.5]$, and so on. Given the $\mu(X_1)$, we derive the fuzzy unidimensional indicator $I(X_1)$ as in (1), and we compare the results related to (5) and to (6), thus investigating the effects of $\mu(X)$ on the poverty indicator.

Table 1 reports in the first column the head count ratio, equal to 0.202, in the second the fuzzy indicator based on (5), equal to 0.264, while in the following columns are summarized the results of 100000 random extractions of $\mu(X_1)$ as in (6). We report, in the first row some values of $I(X_1)$, ranging from 0.097 to 0.375, and from the third to the fifth row the mean of the respective simulated $\mu(X)$. It is possible to observe how extreme values of $I(X_1)$ occur only in correspondence with unlikely values of $\mu(X)$: $I(X_1) = 0.097$, for example, requires $\mu_{11} = 0.454$, which is clearly

an inadmissible value. When, however, $\mu(X)$ remains within acceptable ranges, the index $I(X_1)$ does not show substantial variations.

Table 1 Head count ratio with $\mu(X)=(2)$, fuzzy poverty indicator with $\mu(X)=(5)$ and fuzzy poverty indicator with simulated $\mu(X)=(6)$, Italian households 2016

$I(X_1)$	0.202	0.264	0.097	0.126	0.166	0.186	0.206	0.226	0.306	0.375
$\mu(X_1)$	(2)	(5)	(6)	(6)	(6)	(6)	(6)	(6)	(6)	(6)
μ_{11}	1.00	1.00	0.454	0.610	0.747	0.792	0.822	0.844	0.866	0.914
μ_{12}	1.00	0.90	0.276	0.340	0.450	0.516	0.559	0.591	0.626	0.683
μ_{13}	0	0.50	0.133	0.163	0.212	0.243	0.284	0.329	0.389	0.443
μ_{14}	0	0	0.009	0.015	0.026	0.033	0.043	0.054	0.155	0.233

The second inequality factor X_2 which we introduce is the educational level of the head of the household. The unidimensional fuzzy indicator which we propose is based on $k = 3$ classes and on the following indicator function:

$$\mu(X_2) = \begin{cases} 1.0 & \text{if } x_{2i} = \text{none} \\ 0.9 & \text{if } x_{2i} = \text{elementary} \\ 0 & \text{if } x_{2i} = \text{other} \end{cases} \quad (7)$$

As for X_1 , also for the educational level we develop a simulation study, randomly extracting $\mu(X_2)$ from an uniform distribution on the basis of the following scheme:

$$\mu(X_2) = \begin{cases} \mu_{21} & \text{with } \mu_{21} \in [\mu_{22} + 0.1, 1] \\ \mu_{22} & \text{with } \mu_{22} \in [\mu_{23} + 0.1, 0.66] \\ \mu_{23} & \text{with } \mu_{23} \in [0, 0.33] \end{cases} \quad (8)$$

We calculate $I(X_2)$ as in (1), and in Table 2 we compare the results obtained by using (7) and 100000 random extractions of (8). As for $\mu(X_1)$, also the simulated $\mu(X_2)$ suggest that for the index $I(X_2)$ anomalous results are observed only in correspondence with anomalous values of $\mu(X_2)$.

Table 2 Unidimensional poverty indicator (educational level of the head of the household, Italy 2016): simulated $\mu(X) = (8)$ and exogenous $\mu(X) = (7)$

$I(X_2)$	0.056	0.106	0.155	0.206	0.245	0.275	0.305	0.404	0.183
$\mu(X_2)$	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(7)
μ_{21}	0.564	0.693	0.763	0.782	0.787	0.797	0.803	0.904	1.00
μ_{22}	0.161	0.299	0.431	0.458	0.478	0.493	0.510	0.641	0.90
μ_{23}	0.010	0.035	0.064	0.121	0.167	0.202	0.236	0.326	0.50

The multidimensional fuzzy poverty index I , obtained as (3), is reported in Table 3, together with the unidimensional indices $I(X_1)$ and $I(X_2)$ and the weights w_1 and w_2 . The last column of the table shows the index calculated on (6) and (7), while in

Table 3 Multidimensional poverty indicator I (equivalent income X_1 and educational level of the head of the household X_2 , Italy 2016): simulated $\mu(X) = (6)(8)$ and exogenous $\mu(X) = (5)(7)$

I	0.089	0.146	0.186	0.206	0.236	0.265	0.305	0.393	0.219
$\mu(X)$	(6)(8)	(6)(8)	(6)(8)	(6)(8)	(6)(8)	(6)(8)	(6)(8)	(6)(8)	(5)(7)
$I(X_1)$	0.178	0.241	0.253	0.260	0.265	0.276	0.305	0.389	0.264
$I(X_2)$	0.044	0.097	0.149	0.178	0.226	0.270	0.314	0.398	0.183
w_1	0.763	0.634	0.610	0.598	0.587	0.567	0.519	0.411	0.578
w_2	1.378	1.026	0.839	0.763	0.658	0.577	0.507	0.400	0.736

the first columns some of the values obtained from the simulation are summarized. By comparing the last column with the simulated results we confirm also for the index I a strong stability for reasonable values of $\mu(X_j)$ and of $I(X_j)$.

Our Monte Carlo study also allows us to investigate many other aspects of fuzzy methods, such as the role of h and the values of k , which are not reported here for space constraints.

4 Conclusions

The robustness of fuzzy indicators for poverty analysis is analysed through a Monte Carlo study, with particular emphasis on the issues related to the choice of $\mu(X)$ and their effects on the index. Even in the absence of constraints on the first and last value of $\mu(X)$ (which usually are, respectively, equal to 1 and 0), the simulated values suggest an extremely satisfactory robustness of the fuzzy poverty indicators, thus overcoming the objection related to the subjectivity of the choice of $\mu(X)$. Furthermore, the simulation study proves to be a particularly promising tool to investigate further characteristics of the multidimensional fuzzy poverty indicators.

References

1. Betti, G., Cheli, B., Gambini, R.: A statistical model for the dynamics between two fuzzy states: theory and an application to poverty analysis. *Metron*, **62**, 391–411 (2004)
2. Cerioli, A., Zani, S.: A fuzzy approach to the measurement of poverty. In Dagum, C., Xenga, M.: Income and wealth distribution, inequality and poverty. Springer, Berlin, 272–284 (1990)
3. Cheli, B., Lemmi, A.: A totally fuzzy and relative approach to the multidimensional analysis of poverty. *Economic Notes*, **24**, 115–134 (1995)
4. Dagum, C., Costa, M.: Analysis and measurement of poverty. Univariate and multivariate approaches and their policy implications. In Dagum, C., Ferrari, G.: Household Behaviour, Equivalence Scales, Welfare and Poverty. Springer, Berlin, 221–271 (2004)
5. Haagenars, A.J.M.: The perception of poverty. North Holland, Amsterdam (1986)
6. Lemmi, A., Betti, G.: Fuzzy set approach to multidimensional poverty measurement, Springer, Berlin (2006)
7. Sen, A.K.: Inequality rexamined. Harvard University Press, Cambridge (MA) (1992)

Text Based Pricing Modelling: an Application to the Fashion Industry

Modellazione dei prezzi basata su dati testuali: un'applicazione all'industria fashion

Federico Crescenzi, Marzia Freo and Alessandra Luati

Abstract By using internet as a source of data, we estimate pricing models based on the information contained in the description of items on sale. The novel application is on one category of the Fashion industry. Our estimation strategy uses text mining and methods of sparse modelling, namely shrinkage methods and dimension reduction methods, with the aim of obtaining a model which gives the best out-of-sample predictive performance with a high level of interpretability.

The results show that compared with the simple predictor, the average price, the models developed in the paper produce a decrease in the pricing error which is up to 7.7% when the brand is considered and up to 58.3%, when the brand is not considered, in the case when shrinkage methods are used. When dimension reduction methods are used, the decrease is up to 41.4% when brand is included and up to 66.8% in the no-brand case.

Abstract *L'obiettivo del presente lavoro è di stimare modelli di prezzo usando Internet come fonte di dati, cioè sfruttando l'informazione contenuta nella descrizione degli articoli in vendita. La strategia di stima utilizza metodologie di text-mining, metodi di modellazione di shrinkage, metodi per la riduzione della dimensione, e loro combinazioni, con l'obiettivo di identificare il modello con maggiore performance predittiva. L'applicazione originale è sull'industria della moda. I principali risultati empirici mostrano che i modelli presentati riducono l'errore di previsione del prezzo rispetto al predittore base, prezzo medio, nel caso di metodi di selezione delle variabili fino al 7.7%, quando si considera la marca, e fino al 58.3%, quando non si considera. La riduzione è fino al 41.4% quando si considera la marca, e al 66.8%, quando non si considera, nel caso di metodi di riduzione delle dimensioni.*

Key words: Text Mining, Sparse Modelling, Online Prices, Fashion Industry.

Introduction

The aim of this paper is to develop a predictive model for prices of fashion items, based on their description, available on e-commerce platforms. The rationale is that the most relevant features of an item are summarised in the few words used to describe it. This problem is of relevant interest both on the applied side, due to the growing worldwide role of ecommerce, and on the methodological side, as new data and methods, say new smart approaches, are required to face problems which were previously afforded based on traditional methodologies, such as hedonic evaluation and its impact on price index theory or the specification of pricing algorithms.

For the purposes of the present study, the available information for each product is summarised by the following items: price, category, producing brand, and a textual field which contains a concise description of the item. With the aim of solving a prediction problem based on such information, we combine two main fields of statistics: text analysis and sparse modelling.

As far as text analysis is concerned, there is a consolidated literature on the use of text data for clustering and classification (Berry and Kogan 2010). On the other hand, only few contributions are concerned with textual data processed through text mining procedures and used as predictors of quantitative responses. Some of them focus on price prediction in financial markets, where texts are generally processed by sentiment analysis. Other contributions are concerned with the use of texts to extract relevant attributes to explain prices in real estate (Nowak and Smith, 2017; Foster, Liberman and Stine, 2013). Recently, Mercari, Japan's biggest community-powered shopping app, launched a Kaggle competition to identify an algorithm that automatically deliver product prices to online retailers. The reaction to the competition was immediate, with several machine learning models among the competitors (<https://www.kaggle.com/c/mercari-price-suggestion-challenge>).

As far as sparse modelling is concerned, despite the short description attached to each item, after the text-processing phase, the set composed by the union of words contained in all the considered descriptions, is generally huge. As each word is a feature or a potential predictor, estimation requires methods for large-dimensional parameter spaces. In situations where the number of predictors p is large relative to the number of observations n , the literature proposes two main ways of restricting the attention to lower dimensional subspace: shrinkage methods and dimension reduction methods (see Tibshirani, 1996). Both methods have been applied in text regression: the Lasso penalized regression or its modified version in Nowak and Smith (2017) are attributable to the class of shrinkage methods; the Singular Value Decomposition used by Foster, Liberman and Stine (2013) and the Latent Semantic Indexing (Deerwester et al. (1990), belong to the class of dimension reduction methods. Another stream of the literature has faced the problem of pricing by means non parametric methods, such as an extension of neural network and recursive trees as applied in Mercari.

It is often the case than the results obtained by these methods provide results that may be hard to interpret, while we have the twofold aim to predict and to explain prices in function of textual features. Interpretability of the results is a key issue for

the scope of the paper, hence, between competing models having similar predictive power, we shall favour the ones that provide results of easier interpretation.

One relevant point of our analysis is that we aim at predicting price without using the brand name, i.e. based only on product characteristics. The latter may seem an unfavourable condition, due to the obvious relevance of a well-known brand on the price of an item. However, we believe that an analysis performed in the no-brand name can be extended to different datasets. We shall however consider models that include the brand as well as no-brand models.

In section 2, we describe the data and illustrate the text mining procedure adopted. Section 3 discusses the methods and reports the main results, over the trousers category of the section women apparel of each website. Conclusions and some directions for future research are in Section 4.

Data and preprocessing

We use data scraped from Italian e-commerce websites of five fashion brands, from October 2015 to January 2016, on a weekly basis. The brands are chosen to be heterogeneous with respect to the average price of items on sales: two of them are large fast-fashion chains selling at very low prices (from here in afterwards F1, F2), the remaining three found their business on design of enhanced fashion item (from here in afterwards D1, D2, D3). Fast fashion and enhanced design brands are discussed in Cachon and Swinney (date). Each record collects information on price, category, brand and a field of description. It is worth mentioning that text data are in Italian.

After removing duplicated items, we run text pre-processing actions according to the common procedure when dealing with text data. First, we remove punctuation and substitute accented letters with non-accented. We also remove numbers and special characters. The second step is to eliminate the most common stopwords. Stopwords are all those words which are natural in language that do not provide any semantic meaning (i.e. the counterpart of the English words “a”, “of”, “with” etc.). We add to this list words like model, size, cm, wears, which we believe not to add any useful information to our analysis. The final step is stemming, which consists in reducing each word to its root. Lastly, we merge synonyms and words with similar meaning.

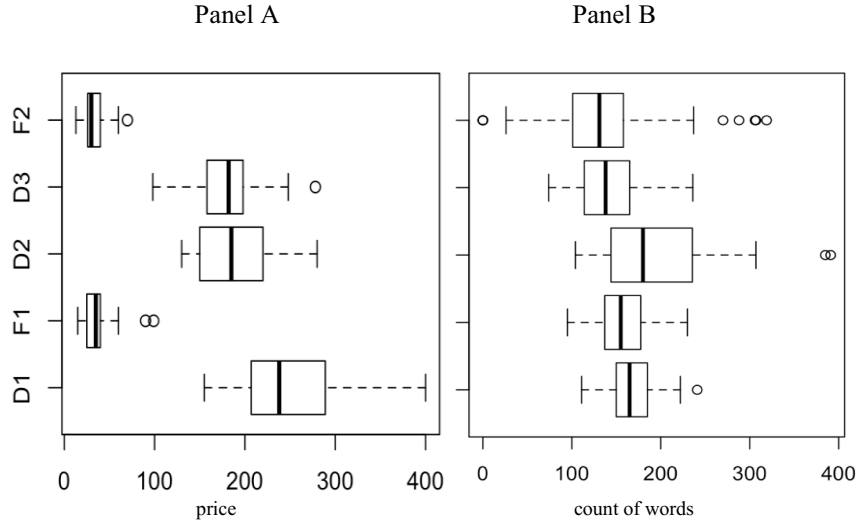
After pre-processing, with the bag of words approach, the number words that we obtain has been reduced of more than 40%, for a total of 682 items and 306 features.

Assuming that each word represents a relevant feature in explaining the price of the item, our purpose is to identify a function which relates prices to features. The main challenge derives from the large number of potential predictors relatively to the number of observations.

On average, retailers display different mean prices (Figure 1, Panel A), though some retailers have similar price distributions. Two couples of fast fashion retailers (F1 and F2) and (D2, D3) have similar, nearly overlapping, price distributions while D1 stands alone from the others. Figure 1, Panel B displays the distributions of word counts by brand that are used to describe each item in the collection of trousers. Few

differences outstand between retailers: some distributions are nearly identical, while D2 and F2 display noticeable variability. On average, approximately 146 words are used to provide descriptions to items.

Figure 1. Distributions of price (left) and count of words (right) by brand. Trousers category.



Methodology and Main Results

The data that we obtain after pre-processing is a collection $\{(y_i, \text{text}_i)\}_{i=1}^N$ where y_i is the price that we observe for item i and text_i is a text-string providing the concise description to the item. The Document Term Matrix, obtained with the standard bag of words approach, counts the words that occur in each item-description; it is a 682×306 dimensional highly sparse matrix. To carry on the analysis, we consider a selection of methods belonging to the two main classes of methods of solving sparse modelling, namely shrinkage methods and dimension reduction methods.

As a shrinkage method, we use the Lasso with different values of tuning parameter. We also find the optimal tuning parameter by minimising the cross-validated mean squared error.

Within the class of dimension reduction methods, we adopt the Partial Least Squares (PLS) and Latent Semantic Indexing (LSI) methodologies. LSI is a popular technique in text mining literature, based on the Singular Value Decomposition to construct a low rank approximation to the Document-Term Matrix. As a new collection of covariates the leading K left singular vectors are considered and are usually given the interpretation of topics.

We estimate a pool of models. The main results are reported in Table 1. The Base fit is the predictor sample mean in the no-brand case, and the sample mean by brand for the brand case. LR stands for Linear Regression using all the features. We name Lasso the Lasso regularized linear regression attained by minimizing the cross-validation Root Mean Squared Error (RMSE), while Lasso-1SE uses the one standard error rule to pick the value of regularization parameter; i.e. the most parsimonious model whose error is no more than one standard error above the error of the best model. This indicates that this more parsimonious model is not statistically worse than the standard Lasso. LSI-K indicates the regression over the first K left singular vectors. LSI-Lasso is attained by selecting left singular values among the first K using the Lasso type regularization that achieves the lowest RMSE. Then in PLS specifications, prices are regressed over the first K partial least squares directions.

Once the different models have been estimated, losses in the out-of-sample performance are computed by using the leave-one-out cross-validation approach. Losses are computed using the Root Mean Squared Error. For ease of interpretation, we compute the Root Relative Squared Error (RRSE), which normalizes the RMSE of any model to the one of the specific base predictor.

Table 1. Results of cross-validated prediction by specification

TROUSERS	<i>Without brand</i>			<i>With brand</i>		
	Nr. Covariates	RMSE	RRSE	Nr. Covariates	RMSE	RRSE
Base	0	83.339	1.000	4	26.707	1.000
Linear Regression	306	117.805	1.414	310	87.215	3.266
Lasso	133	34.737	0.417	81	24.662	0.923
Lasso-1SE	67	36.660	0.440	35	24.850	0.930
LSI-K	100	30.412	0.365	90	22.790	0.853
LSI- Lasso	89	29.433	0.353	77	22.481	0.842
PLS	1	37.053	0.445	5	25.366	0.950
PLS	4	27.629	0.332	8	15.653	0.586

Specifications without brand always achieve prediction errors higher than the ones attained by models with brand; clearly, the brand is the most important feature in determining prices. Linear regression estimations obtained using all the features largely increase the out-of-sample RMSE, because of the overfitting and the sparsity of the design matrix. Adding a Lasso type regularization improves out-of-sample prediction as compared as both the full text-and the base regressions. For the no-brand case, the RMSE drops to 34.737 (approximately 58.3% lower than for the base regression). Regarding the brand case, the estimated RMSE drops to 24.662, i.e. a 7.7% reduction as compared to the regression only over brands. Regression on K = 90 left singular vectors and brand dummies attains the minimum point estimate of RMSE at 22.790. This slightly improves the Lasso prediction, as one observes a 8.6% reduction. The same tuning procedure run by omitting the retailer dummies yields RMSE=30.412 but it needs K = 100 left singular vectors to achieve the minimum. Again, this improves the Lasso prediction by 12.5%. Price regression on K = 4 partial least squares directions yields the best prediction results. For a regression without brand dummies, we get RMSE=27.629. Plugging these directions in a OLS model

with brand dummies we get $RMSE = 15.653$. As a whole, while the Lasso-1SE reduces the prediction error of 56% and 7% respectively in the no-brand and brand case, PLS improves Lasso performance and reduces the RMSE of 66.8%, and 41.4%.

Conclusions

We have provided a comparison of text based predictive models for prices in the fashion system, by using internet as a source of data. With such approach, we are able to control for characteristics easy to observe, like item compositions in terms of materials, and to mitigate the bias due to omitted variables, by incorporating in the model covariates difficult to measure such as fit or appealing design. Due to high dimensionality, least squares estimation is prone to overfitting. Regularisation or dimensionality reduction techniques are employed to minimize pricing errors. Empirically, both approaches strongly appeared to reduce out-of-sampling error with respect to linear regression. Dimensionality reduction methods, LSI, and also more PLS provide better performance than Penalized regression. This happens both when the brand variable is and is not included among covariates.

References

1. Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: applications and theory*. John Wiley & Sons.
2. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391-407.
3. Foster, D. P., Liberman, M., & Stine, R. A. (2013). Featurizing text: Converting text into predictors for regression analysis. The Wharton School of the University of Pennsylvania, Philadelphia, PA.
4. Nowak, A., & Smith, P. (2017). Textual analysis in real estate. *Journal of Applied Econometrics*, 32(4), 896-918.
5. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 267-288..x

Model based clustering in group life insurance via Bayesian nonparametric mixtures

Raggruppamento basato sul modello nel settore assicurativo: un approccio bayesiano nonparametrico

Laura D'Angelo

Abstract Experience rating allows insurance companies to adjust the premium for a certain policyholder on the basis of the loss experienced by similar insured parties. In group life insurance, because of the lack of complete information, clustering policyholders is particularly challenging. To address this issue we adopt a model-based clustering approach using flexible Bayesian nonparametric mixtures, which take into account the discrete nature of data, consisting of claim counts. We consider a Pitman-Yor mixture of Rounded Gaussian kernels, which provides more flexible and robust results than standard Poisson mixtures. We show how this approach leads to more accurate inference compared to standard Dirichlet process mixtures of Poisson, through an application to data arising from a portfolio of groups of workers.

Abstract *Nell'ambito delle assicurazioni sulla vita collettive è di particolare importanza definire il rischio associato ad un determinato gruppo: ciò viene svolto sulla base del numero di richieste di risarcimento osservate per gruppi con caratteristiche simili. Tuttavia, in questo contesto, individuare i gruppi simili risulta particolarmente complesso a causa della presenza di una grande eterogeneità non osservata. Per rispondere a questo problema, di seguito è proposto un approccio Bayesiano nonparametrico basato su un modello mistura. In particolare, per tenere conto della natura discreta dei dati, si propone una mistura di kernel Gaussiane arrotondate basata sul processo di Pitman-Yor. Attraverso un'applicazione a un portafoglio assicurativo, si mostra come questo modello fornisca risultati più accurati rispetto a una mistura basata sul kernel Poisson.*

Key words: Pitman-Yor Process, Rounded Gaussian, Dirichlet Process, Poisson mixtures

Laura D'Angelo
Department of Statistical Sciences, University of Padova, Via C. Battisti 241, Padova, Italy, e-mail:
laura.dangelo.1@phd.unipd.it

1 Introduction

An accurate assessment of the risk associated to policyholders is a fundamental task in group life insurance. This is usually achieved through experience rating, which allows to adjust the premium for a certain policyholder on the basis of the loss experienced by similar insured parties. While in individual life insurance one can record many relevant factors affecting risk, in group life insurance clustering observations becomes a tough challenge because of the presence of many unrecorded risk factors, which lead to a great unobserved heterogeneity.

Our study is motivated by the analysis of data of a real group life insurance portfolio, which consists of the claim counts and total risk exposure of 72 groups of workers insured in the period 1982-1985. The data are also analyzed in [7], [6] and [1]. In their work, [1] assert how standard parametric models do not allow to adequately describe the heterogeneity between groups, and adopt a more flexible approach using Bayesian nonparametric mixtures: specifically, they propose to use a Dirichlet Process mixture of Poisson kernels. Although apparently flexible, a deeper analysis reveals how this model has some drawbacks, arising from both choices of the kernel and of the mixing measure. In fact, even if the Poisson kernel is a natural choice in the context of count data, it lacks flexibility, having a single parameter controlling both the mean and the variance, and forcing them to be equal. This lack of flexibility is extended also to nonparametric mixtures: for example, all distributions which are under-dispersed can not be consistently estimated [2, 3]. We propose to use a Rounded Gaussian kernel, a more flexible kernel for count data introduced in [2], based on rounding of continuous kernels.

Regarding the mixing measure, even if the Dirichlet Process [4, 5] is a popular nonparametric prior, it does not allow to have full control of the clustering structure and the results are indeed heavily affected by the prior choice of the concentration parameter. Using a more general prior can relieve this problem: specifically, we adopt a Pitman-Yor process [8], a tractable generalization of the Dirichlet Process, which introduces one further parameter. In their work, [3] show how also in the case of count data, using this process as a prior in nonparametric mixtures can lead to more robust inference than the Dirichlet process, as for varying prior centering one obtains more stable results.

2 Model

Let $Y_i \in \mathcal{Y}$ be the discrete random variable representing the claim count for group i , $i = 1, \dots, n$. We propose to model the distribution $p(y)$ as a nonparametric mixture

$$p(y) = \int K(y|\psi) dP(\psi),$$

where $K(y|\psi)$ is a kernel on (\mathcal{Y}, Ψ) and Ψ is a parametric space. As a prior for the unknown mixing measure P , we assume a Pitman-Yor process, $P \sim PY(\sigma, \theta, P_0)$, which includes the Dirichlet Process for $\sigma = 0$.

Concerning the choice of the kernel $K(y|\psi)$, we adopt a flexible Rounded Gaussian (RG) kernel. Briefly, the idea is to consider an underlying continuous variable $Y^* \in \mathcal{Y}^*$ with density f : the probability mass function of a discrete variable Y can be obtained through a set of thresholds on the support of Y^* as

$$p(j) = \int_{a_j}^{a_{j+1}} f(y^*) dy^*$$

where $\{a_j\}_{j=0,1,\dots,+\infty}$ is a predefined sequence of thresholds such that $a_0 = \min\{y^* : y^* \in \mathcal{Y}^*\}$ and $a_{+\infty} = \max\{y^* : y^* \in \mathcal{Y}^*\}$. For the underlying continuous kernel, we assumed a Gaussian distribution, with thresholds $\{a_j\}_{j=0}^{+\infty} = \{-\infty, 0, 1, 2, \dots, +\infty\}$. Indicating with μ and τ respectively the location and precision parameter of the rounded Gaussian kernel, $RG(\mu, \tau^{-1})$, the resulting model can be expressed through its hierarchical representation as:

$$\begin{aligned} Y_i | \mu_i, \tau_i &\sim RG(\mu_i, \tau_i^{-1}) \\ (\mu_i, \tau_i) | P &\sim P \\ P &\sim PY(\sigma, \theta, P_0) \end{aligned} \quad (1)$$

For simplicity of computation we chose a conjugate base measure: $P_0(\mu, \tau) = N(\mu; \mu_0, \kappa \tau^{-1}) \text{Gamma}(\tau; \alpha, \beta)$, where the parameters μ_0 and κ were fixed equal to the sample mean and variance, respectively.

3 Application to insurance data

The dataset consists of the number of claims and risk exposure for 72 groups of workers, where the risk exposure is computed as the total exposure for all individuals in the group. Figure 1 shows the histogram of the number of claims normalized by the corresponding exposure: for many groups the number of deaths results to be zero, regardless of the amount of exposure; most of the remaining groups concentrate around small values, but there are few isolated groups for which the number of claims is definitely high: it seems indeed reasonable to assume a non-homogeneous portfolio.

On these data we computed the posterior under model (1), where the exposure was introduced as a multiplicative factor on the mean of the latent variable. In this way, the expected number of deaths is proportional to the duration of the exposure. On the contrary, we assumed no restrictions for the variance and it is independent of the exposure: the possibility to distinguish its effect on the mean and on the variance is another advantage of the Rounded Gaussian kernel, as for the Poisson kernel the only way is to constrain both the mean and the variance to be equally affected.

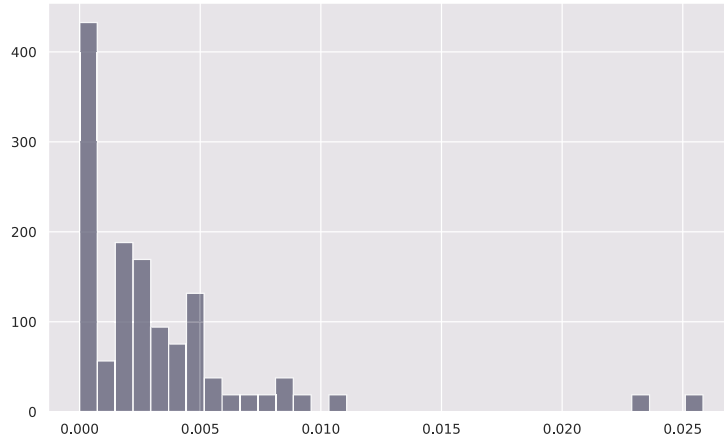


Fig. 1 Histogram of the number of deaths per unit of exposure.

Concerning the prior specification, the parameter σ of the Pitman-Yor process was fixed equal to 0.6, as positive values of this parameter lead to more robust results in the clustering structure of the mixture. This behavior is well displayed in [3], where through a simulation study the authors show how adopting a Dirichlet process (which corresponds to $\sigma = 0$) causes the posterior estimates of the number of clusters to be heavily affected by the prior expectation, while, for increasing σ , the posterior mean stabilizes even for different prior specifications. Concerning the parameter θ , we fixed it conditionally to the sample size and σ to have a prior expected number of clusters equal to 15. This prior specification also allows us to obtain a better comparison between the results of our model and of a Dirichlet Process mixture of Poisson kernels, as we are able to center both processes on the same prior expectation.

The posterior mean probability mass function (pmf) obtained with both mixtures, with the exposure fixed equal to three different levels, are displayed in Figures 2 and 3. At a first glance, the pmfs from the two models look very different, however, analyzing the resulting distributions for the same level of exposure, we can observe some similarities. In fact, the location of the mixture components is similar for both models: for an exposure level $E = 1000$, both pmf have maximum in 2 and, in general, most probability mass is assigned to values between 0 and 10. When the exposure is set to 5000 and 10000, the differences increase, as only for the components closest to zero we can identify clearly the same location. In fact, for the Poisson mixture, the components away from zero are less clear, as they are spread because of the forced increased variance, which is itself multiplied by a factor equal the exposure. This behavior is clearly enhanced for an exposure equal to 10000, as the components further from zero are completely flattened. This drawback does

not exist for the RG mixture and, indeed, all mixture components are well defined for every level of exposure. Moreover, as it is reasonable to guess, higher levels of exposure shift the location of the distribution to higher values of claim counts.

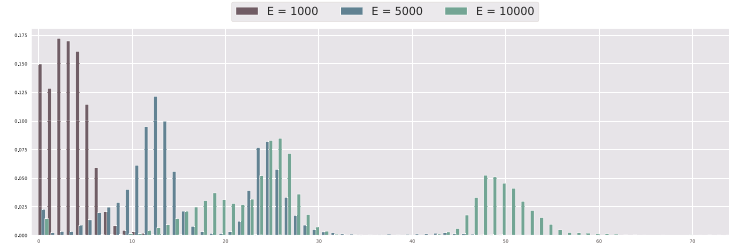


Fig. 2 Estimated pmf using a PYP mixture of RG kernels. Colors correspond to levels of exposure.

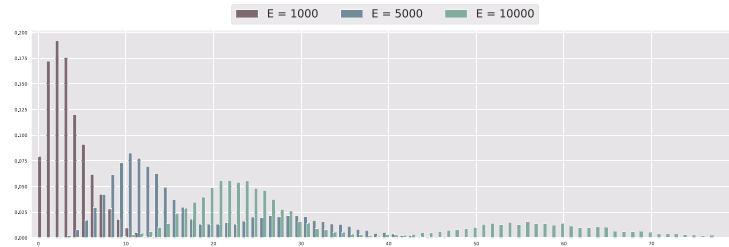


Fig. 3 Estimated pmf using a DP mixture of Poisson kernels. Colors correspond to levels of exposure.

4 Conclusion

In this application we analyzed a group life insurance portfolio, where for each group the number of deaths and the overall exposure were recorded. We showed how the heterogeneity between groups required a nonparametric approach and, to address this issue, we proposed a flexible Pitman-Yor process mixture of Rounded Gaussian kernels. To compare this model with a more “classical” approach, we also estimated a Dirichlet Process mixture of Poisson kernels, although some previous results suggested the inadequacy of this model under many conditions. The application of these mixtures to our data further supported the use of the Rounded Gaussian kernel compared to the Poisson kernel, as the constraints on the variance of the latter led to unlikely flattened and over-dispersed mixture components.

References

1. Brown G.O., Buckley W.S.: Experience rating with Poisson mixtures, *Annals of Actuarial Science* **2** 304321 (2015)
2. Canale A., Dunson D.: Bayesian Kernel Mixtures for Counts. *J. Am. Stat. Assoc.* **106** 1528–1539 (2011)
3. Canale A., Pruenster I.: Robustifying Bayesian nonparametric mixtures for count data. *Biometrics* **73** 174–184 (2017)
4. Ferguson T.S.: A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist.* **1** 2 209–230 (1973)
5. Ferguson T.S.: Prior Distributions on Spaces of Probability Measures. *Ann. Statist.* **2** 4 615–629 (1974)
6. Haastrup, S.: Comparison of Some Bayesian Analyses of Heterogeneity in Group Life Insurance. *Scand. Actuar. J.* **1** 2–16 (2000)
7. Norberg, R.: Experience Rating in Group Life Insurance. *Scand. Actuar. J.* **4**, 194–224 (1989)
8. Pitman J., Yor M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25** 2 855–900 (1997)

Smart Tools for Academic Submission Decisions: Waiting Times Modeling

Strumenti "Smart" per sottoporre i manoscritti accademici: modelli per i tempi di attesa

Francesca De Battisti - Giancarlo Manzi

Abstract This paper illustrates the results of a multilevel analysis aiming at the determinants of peer-review waiting times until acceptance across some of the top statistics & probability journals. Classical multilevel tools are used for analyzing waiting times until acceptance for around 3,500 articles. Results reveal the importance of the number of authors, their academic level and nationality, together with the quality level of the journal as the most important factors affecting waiting times.

Abstract Questo articolo presenta i risultati di un'analisi multilivello sui fattori che determinano il tempo di attesa per il processo di peer review in alcune delle principali riviste accademiche di statistica e probabilità. Strumenti classici della metodologia multilivello vengono utilizzati per analizzare i tempi di attesa fino all'accettazione di circa 3.500 articoli. I risultati rivelano che il numero degli autori, il loro livello accademico, la loro nazionalità e il livello della rivista sono i fattori più importanti che influenzano i tempi di attesa.

Key words: Peer review, Statistical journals, Waiting times, Hierarchical models

1 Introduction

In recent years academic peer-review waiting times are constantly increasing. Figure 1 shows the yearly average number of days between submission and acceptance of published articles in the Advances in Data Analysis and Classification (ADAC)

Francesca De Battisti

Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Via Conservatorio 7, 20122 Milan, Italy. e-mail: francesca.debattisti@unimi.it

Giancarlo Manzi

Department of Economics, Management and Quantitative Methods and Data Science Research Center, Università degli Studi di Milano, Via Conservatorio 7, 20122 Milan, Italy. e-mail: giancarlo.manzi@unimi.it

journal¹ between 2011 and 2016, and depicts a scenario where it has more than doubled from 224 days in 2013 to 474 days in 2016. This is a situation that not only ADAC but also other journals are experiencing, due to lack of peer reviewers, increasing number of submissions, among other reasons.

The scientific peer review process has always been labour intensive, costly, and often slow, with resulting delays in publication [9]. Nowadays we are witnessing the progressive introduction of stricter rules and guidelines by editors and publishers with the latter increasingly favoring pay-to-publish process, with the side effect of a general value downsizing of academic research [12].

As for the impact of academic research, still the impact factor (IF) of academic journals is considered one of the cornerstones for the evaluation of journals [1, 11], whereas for the personal reputation of authors, citation-based indexes like the Hirsh index *does seem to be able to identify good scientists* [2, 8]. However, they both have been severely criticized and are also regarded as not being able to reflect a broader impact of research [3, 7, 10]. Notwithstanding this drawbacks, they reflect the value of papers and can be referred as important measures to evaluate their quality.

The structure *publisher*→*journal*→*article* might in truth be viewed as a multi-level structure and therefore exploited to reveal hidden characteristics of the scientific research production. Therefore, this paper illustrates the results from multiple multilevel models focusing on finding determinants affecting peer-review waiting times until acceptance across around 3,500 articles published between 2011 and 2016 in 8 statistics & probability journals (chosen among the top 38 journals in

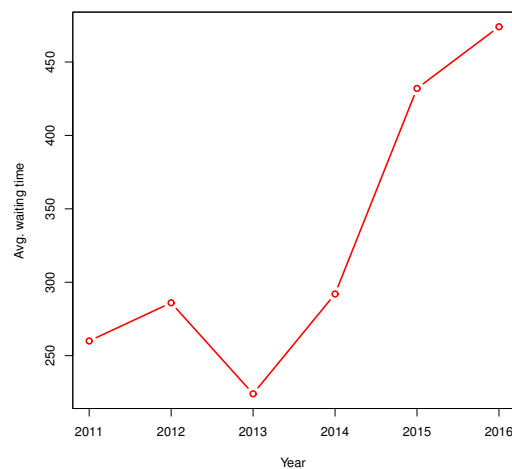


Fig. 1 Average waiting times for articles published in ADAC - Years 2011-2016
(Source: Author's calculations based on Springer ADAC website)

¹ ADAC was ranked 31st out of 123 journals in the Web of Science Journal Citation Report Impact Factor 2017 ranking in the Statistics & Probability category.

the 2017 Thomson Web of Science Statistics & Probability ranking) from different publishers. The 2-level structure *journal*→*article* is chosen for this analysis.

2 Journal, variable and article selection

In constructing our data set, we aimed at top statistical journals. We chose 8 journals among the top 40 statistical journals ranked in the Statistics & Probability category of the 2017 Thomson Journal Citation Report published by multiple publishers. The chosen journals were the *Journal of Statistical Software* (J Stat Softw - ranked 1st), *Fuzzy Sets and Systems* (Fussy Set Syst - ranked 8th), the *Journal of the Royal Statistical Society - Series A - Statistics in Society* (JRSSA - ranked 11th), the *Journal of the American Statistical Association* (JASA - ranked 16th), the *Annals of Probability* (Ann Prob - ranked 19th), the *Journal of Business Economic Statistics* (JBES - ranked 20th), *Advanced in Data Analysis and Classification* (ADAC - ranked 31st), and *Biostatistics* (ranked 38th). This selection was done in order to cover multiple fields of statistics and probability, and allow for multiple regional covering, considering both data analysis journals and theoretical journals.

In selecting potential predictors and response variables for our models, we used both manual and automatic web scraping and data parsing on the journals' websites. For each article (uniquely identified by the Document Object Identifier - DOI) we extracted the date of submission (always available) as the date of the "birth" of an article in the peer-reviewing process, and, as the date of the "death" of an article, the date of acceptance when this was available, otherwise the date of online publication or the date of final revision. The difference between the date of acceptance and the dates of online publication or the final revision might be considered negligible with respect to the average waiting times for acceptance. It ranged between a few days and one or two months.

We also searched for the keyword "Bayes" in the articles and classified them as "Bayesian articles" (value equal to 1) if it was clear from the occurrences that the Bayesian method was used extensively in the article (variable *BAYES*). We operationalized a variable indicating whether the article was openly accessible or not (*OPEN*). From the Scopus database (www.scopus.com) we extracted the number of citations and the *h* index for each author. Computed variables entering the models were the length in days of the waiting times between submission and acceptance (*AGE*), the average *h* index among authors (*AVG_HI*), the standard deviation of the *h* index of the authors (*SD_HI*), the average monthly number of Scopus citations per article (*MONTH_SCOP_CIT*), the number of authors of the articles (*NUMBER_AUTHORS*), two dichotomous variables representing respectively the presence of young or non-expert researchers among authors (expressed by a Scopus *h* index lower than 5 - variable *JUNIOR_LESS5*), and the presence of expert researchers (expressed by a Scopus *h* index greater than 20 - variable *SENIOR_MORE20*), a dichotomous variable equal to 1 if all the authors were affiliated to an US institute and 0 otherwise (*USA_ALL*), a dichotomous variable equal to 1 if at least one

of the authors was affiliated to an US institute and 0 otherwise (*USA*), a dichotomous variable equal to 1 if all the authors were affiliated to an institute of the same country (*SAME_COUNTRY*), a dichotomous variable equal to 1 if the main author (detected by the highest *h* index among the authors) was affiliated to an institute of the same country of the institute to which the editor(s) in chief of the journal was affiliated (*SAME_NATIONALITY_EDS*), the country of the institute to which each author was affiliated, a categorical variable representing the continent of the institute to which the most expert author belongs. Journal-level variables were the 2017 Thomson Reuters impact factor (*IF*), the journal frequency in days (*PERIODICITY*), the age of the journal since its foundation (*AGE_JOURNAL*), the AI index (i.e. the article influence score measuring the average influence of a journal article over the first five years after publication - variable *AI*).

Articles considered in the analysis were only non-solicited research articles not included in special issues. We excluded presidential addresses, editorials, book reviews, conference proceedings, code snippets, comments, rejoinders and corrections. After this inclusion procedure, a total number of around 3,500 articles formed our database subdivided across journals as reported in Table 1.

Table 1 Distribution of articles across journals

<i>Journal</i>	<i>Publisher</i>	<i>Country</i>	<i>Eds' country</i>	<i>No. of arti- cles in the database</i>	<i>%</i>
JRSSA	Royal Stat.Soc./Wiley	UK	UK	246	7.03
ADAC	Springer	GER	ITA, GER, JAP	127	3.63
JBES	Am.Stat.Ass./ Tay- lor&Francis	USA	USA	252	7.20
Ann Prob	Inst.Math.Stats./ Bernoulli Society	USA	USA	485	13.85
Biostatistics	Oxford Uni. Press	UK	NED, USA	334	9.54
Fuzzy Set Syst	Elsevier	NED	BEL, FRA, GER, SPA	982	28.05
JASA	Am.Stat.Ass./ Tay- lor&Francis	USA	USA	734	20.97
J Stat Softw	UCLA Dept.Stats	USA	AUT, SWI, GER	341	9.74

3 Model used and results

A standard random intercept-random slope multilevel model with two levels (article=level 1; journal=level 2) without interactions between first and second level predictors was used to model the responses "days from submission to acceptance":

$$Y_{ij} = \gamma_{00} + \gamma_{10}\mathbf{X}_{ij} + \gamma_{01}\mathbf{Z}_j + u_{0j} + u_{1j}\mathbf{X}_{ij} + \varepsilon_{ij}, \quad (1)$$

where i is the index for article, j is the index for journal, \mathbf{X} is a vector of level 1 predictors, and \mathbf{Z} is a vector of 2-level predictors.

Table 2 shows the results of model (1) applied to our data set. The dependent variable "time between submission and acceptance" is regressed towards various independent variables. The null model (column I), the model with only 1-level predictors (column II), the full model (column III) and the best model when only AI is retained among 2nd-level predictors (column IV) are shown.

Table 2 Multilevel model (1) applied to the article data set. Dependent variable: time between submission and acceptance. SE in parentheses. Significance: * 0.10; ** 0.05; *** 0.01

Variable	I (null)	II (1-level only)	III (full model)	IV (best model)
Intercept	460.21(57.89)***	452.43(59.58)***	270.07(85.37)***	302.20(25.63)***
First level predictors:				
BAYES		5.26(11.78)	5.27(11.79)	4.79(11.67)
OPEN		1.21(17.39)	-3.03(17.60)	
MONTH_SCOPUS_CIT		-1.46(1.54)	-1.48(1.54)	
AVG_HI		-0.22(0.73)	-0.22(0.73)	-0.93(0.46)**
SD_HI		-0.58(0.92)	-0.59(0.92)	
JUNIOR_LESS5		9.53(9.74)	9.58(9.74)	5.71(8.80)
SENIOR_MORE20		-11.92(11.89)	-12.05(11.89)	
NUMBER_AUTHORS		8.05(3.88)**	8.11(3.89)**	6.00(3.46)*
USA_ALL		0.60(17.16)	-0.05(17.16)	
USA		2.99(14.11)	2.34(14.11)	
SAME_COUNTRY		0.53(11.38)	0.97(11.40)	
SAME_NATIONALITY_EDS		-17.56(10.19)*	-16.92(10.20)*	-18.53(9.32)*
Second level predictors:				
PERIODICITY			-0.28(0.66)	
AGE_JOURNAL			-0.43(0.77)	
IF			-9.33(7.24)	
AI			93.64(19.68)***	66.88(8.59)***
AIC:	48066.8	47275.8	47280.8	48021.9

4 Discussion on results and future work

The contribution of this paper is twofold. Firstly, to our knowledge this is the first time a database focused on the time interval between submission and editorial decision and comprehensive of many article features contained in more than one journal is built. The work by Bornmann and Daniel [4, 5] was focused on the time interval between submission and editorial decision of 1899 communication manuscripts, but only considering one journal, the *Angewandte Chemie International Edition*. Secondly, it represents the basis for further and more appealing analysis to be im-

plemented with more sophisticated statistical tools like the multilevel excess hazard survival model proposed by Charvat et al. [6]. Bayesian hierarchical methods are also suitable for this analysis. Results of our analysis show that to some extent the number of authors forming the research team seems to increase the waiting time for acceptance. This could be interpreted as a wasting time from the side of researchers due to more time needed for meetings, taking decisions etc. If the most important researcher among the authors is of the same nationality of one of the editors in chief the waiting time for acceptance decreases. This could be imputed to the attitude of editors being facilitated in communicating with researchers. Finally, the average academic level of authors is negatively associated with the waiting time for acceptance. Among second level predictor, *AI* seems to be the best: the higher the quality level of the journal the longer the waiting time.

There is a lot of future work to do. Among many things, further levels comprising publishers and authors must be considered, the use of a multilevel survival analysis to be compared with the simple multilevel analysis performed here and a research impact analysis based on the Scopus citations per article.

References

- [1] Archambault, E., Larivire, V. (2009) History of the journal impact factor: Contingencies and consequences. *Scientometrics*, 79(3), 635-649.
- [2] Ball, P. (2007) Achievement index climbs the ranks. *Nature*, 448, 737.
- [3] Bornmann, L., Daniel, H.D. (2008) What do citation counts measure? A review of studies on citing behaviour. *J Doc*, 64, 45-80.
- [4] Bornmann, L., Daniel, H.-D. (2010a) How Long is the Peer Review Process for Journal Manuscripts? A Case Study on Angewandte Chemie International Edition. *CHIMIA*, 64(1-2), 72-77.
- [5] Bornmann, L., Daniel, H.-D. (2010b) The manuscript reviewing process: Empirical research on review requests, review sequences, and decision rules in peer review. *Library & Information Science Research*, 32: 5-12.
- [6] Charvat, H., Remontet, L., Bossard, N., Roche, L., Dejardin, O., Rachet, B., Launoy, G., Belot, A. (2016) A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Stats in Med*, 35, 3066-3084.
- [7] Didegah, F., Bowman, T.D., Holmberg, T. (2018) On the Differences Between Citations and Altimetrics: An Investigation of Factors Driving Altimetrics Versus Citations for Finnish Articles. *J Assoc Inf Sci Tech*, 69(6), 832-843.
- [8] Egghe, L. (2010) The Hirsch index and related impact measures. *Information Science and Technology*, 44(1), 65-114.
- [9] Hames, I. (2007) *Peer Review and Manuscript Management in Scientific Journals. Guidelines for Good Practice*. Malden, MA: Blackwell Publishing.
- [10] Holmberg, K., Didegah, F., Bowman, T.D. (2015) The different meanings and levels of impact of altmetrics. Proceedings of the 11th International Conference on Webometrics, Informetrics and Scientometrics & 16th COLLNET Meeting, 26-28 November, New Delhi, India.
- [11] Seiler, C., Wohlrabe, K. (2014) How robust are journal rankings based on the impact factor? Evidence from the economic sciences *J Informetr*, 8, 904-911.
- [12] Teixeira da Silva, J. A. (2017) Fake peer reviews, fake identities, fake accounts, fake data: beware! *AME Med J*, 2(28), 1-3.

On the Use of Control Variables in PLS-SEM

Sull'Uso delle Variabili di Controllo nei PLS-SEM

Francesca De Battisti and Elena Siletti

Abstract Several authors have recently devoted more attention to the control variables methodological issue. Despite many recommendations to handle these variables more efficiently, good practices are still widely disregarded, and especially this topic has not yet been studied in depth for structural equation models. This paper suggests best research practices for researchers who deal with the use of control variables in partial least squares structural equation models.

Abstract *Recentemente diversi autori hanno dedicato attenzione al problema delle variabili di controllo. Nonostante le molte raccomandazioni suggerite per gestire queste variabili in modo più efficiente, le buone pratiche sono ancora ampiamente ignorate, in particolare questo argomento non è stato ancora approfondito nei modelli ad equazioni strutturali. Questo lavoro propone alcune linee guida a chi deve utilizzare le variabili di controllo nei modelli ad equazioni strutturali con stima dei minimi quadrati parziali.*

Key words: Control variables, Structural equation models, Partial least squares

1 Introduction

Control variables are traditionally considered in causal models to rule out alternative explanations for findings or to reduce error terms and increase statistical power. They are variables that do not change and that can cause or be correlated with the causal variable, mediator and outcome. There are two primary means of controlling variables. The first is control by experimental design, whereby the researcher manipulates the nature of the sample. The second is statistical control, whereby

De Battisti Francesca and Siletti Elena
Department of Economics, Management and Quantitative Methods,
Università degli Studi di Milano, Via Conservatorio, 7 - 20122 Milan, Italy;
e-mail: francesca.debattisti@unimi.it, elena.siletti@unimi.it

the researcher measures relevant variables and includes them in the analysis. This approach mathematically partials the effect of the control from the other variables (Becker, 2005). There is no widespread agreement to handling statistical controls (Spector et al, 2000; Breaugh, 2008). Moreover, Spector and Brannick (2011) warn against the misuse of demographic variables because attention should be focused on the mechanisms that explain relations with demographics rather than on the demographics themselves. Becker et al (2016) recently recommend to handle controls more efficiently, paying attention to variables selection, to the methods for measuring and analysing them, to reporting and interpreting results. Becker (2005) considers the controls in structural equation models (SEM), underlined that this issue deserves further attention, and that authors using SEM need to explain why they are treating control variables as they are.

2 Guidelines on the Use of Control Variables in PLS-SEM

SEM allow to study the relationships among latent, not directly observable, variables. Two types of SEM can be distinguished: covariance- and variance-based models. In variance-based models, linear combinations of observed variables are first created as proxies, and then the parameters are estimated by them. Among variance-based SEM methods, partial least squares (PLS) path modelling (Wold, 1985; Lohmöller, 1989) has been called a “silver bullet” (Hair et al, 2011); it is recently used intensively and it will be discussed below. PLS is an iterative algorithm to estimate the different blocks of the measurement model separately and then, in a second step, to estimate the coefficients of the structural model; the aim is to explain the best residual variance of the latent variables and, potentially, of the manifest variables. PLS-SEM consider a sequence of equations that describe the relationships among key theoretical constructs (i.e. the structural model) and a sequence of equations that show the relations among the latent and manifest variables (i.e. the measure model). The presence of a measurement model alone represents a confirmatory factor analysis; the case of a structural model alone represents a path analysis on observed variables. Since the analysis of controls involves the presence of causal links, in the following discussion our attention will be devoted to the structural model alone, and a report of the possible relationships involving controls is presented. To simplify, the usual multiple regression notation is adopted: X is an independent (predictor/exogenous) variable, Y is a dependent (criterion/endogenous) variable, M is a mediator variable, m is a moderator variable, and C is a control. We remember that all these variables are constructs or latent variables, each of which is related to one or more manifest or measured variables. The simplest causal model has two variables: X and Y ; and it can be represented mathematically with a single equation, or with the correspondent path diagram. When we introduce a control, in the system there is only one equation, but with one more variable on the right. Even with only three variables, the scenario is more complex. The third variable could be

independent or dependent, a mediator, or a moderator. These options are represented by Eq. 1, 2, 3 and 4, respectively.

$$Y = \beta_0 + \beta_{X_1Y}X_1 + \beta_{X_2Y}X_2 + \varepsilon_Y \quad (1)$$

$$Y_1 = \beta_{0_1} + \beta_{XY_1}X + \varepsilon_{Y_1} \quad Y_2 = \beta_{0_2} + \beta_{XY_2}X + \varepsilon_{Y_2} \quad (2)$$

$$Y = \beta_{0_1} + \beta'_{XY}X + \varepsilon_{Y'}; \quad M = \beta_{0_2} + \beta_{XM}X + \varepsilon_M; \quad Y = \beta_{0_3} + \beta_{XY}X + \beta_{MY}M + \varepsilon_Y \quad (3)$$

$$Y = \beta_0 + \beta_{XY}X + \beta_{mY}m + \beta_{m*X}mX + \varepsilon_Y \quad (4)$$

With two independent variables the model is represented by only one equation (Eq.1). With two dependent variables we have a system with two equations (Eq.2), and the control can be introduced in three different ways: only in the first, in the second, or in both equations. Dealing with a mediation, we refer to the mathematical representation proposed by Judd and Kenny (1981), as the system in Eq.3, where in the first equation, β'_{XY} , represents the total effect of variable X on Y ; in the second equation, β_{XM} is the effect of X on mediator M ; in the last equation, β_{XY} is the direct effect of X on Y , and β_{MY} is the effect of M on Y . Notably, with ordinary least squares (OLS) regression the first and third equations are fitted as separate regression models, but in PLS-SEM they are fitted simultaneously. Based on the coefficients, the indirect effect can be computed as the product of the β_{XM} and β_{MY} paths or as the difference between β'_{XY} and β_{XY} (i.e. $\beta'_{XY} - \beta_{XY}$). Furthermore, the proportion mediated can be calculated as $\frac{\beta_{XM}\beta_{MY}}{(\beta_{XM}\beta_{MY} + \beta_{XY})}$, $\frac{\beta_{XM}\beta_{MY}}{\beta'_{XY}}$ or $1 - (\frac{\beta_{XY}}{\beta'_{XY}})$. Traditionally, we refer to this kind of link as *partial mediation*, while if the direct effect, β_{XY} , is not significant the mediation is said to be *full*. In mediation analysis, the latter case, in which the total effect, β'_{XY} , is equal to the indirect effect, $\beta_{XM}\beta_{MY}$, is the most interesting because the link between the dependent and independent variables is significant only through the mediator. In this kind of model, the controls can be expressed in different ways: on only the dependent variable Y , on only the mediator variable M or on both of the variables. The way to deal with controls does not change for partial or full mediation models. With a moderator (Eq.4), we need to take its nature into account. Baron and Kenny (1986) defined four cases of moderation by predictor and moderator scales. In the context of PLS-SEM, the predictor is latent; for this reason, only two cases must be considered, where moderator effects are indicated by the interaction of X and m in explaining Y and are measured by β_{m*X} . With continuous moderators, the global effect of X on Y is, therefore, defined as the sum of the simple effect, β_{XY} , and the product, $\beta_{m*X}m$, since the effect of X on Y depends on the value of m (the product term approach (Chin et al, 2003), see Eq.4); the corresponding model with a control is represented as:

$$Y = \beta_0 + \beta_{XY}X + \beta_{mY}m + \beta_{m*X}mX + \beta_{CY}C + \varepsilon_Y \quad (5)$$

If the moderator is non-metric, or if we categorize a quantitative variable, we apply the multi-group analysis (MGA; Henseler et al (2009)). Following this method, the moderator effect is measured by replicating the analysis on subgroups that differ by moderator level. When we conduct an MGA, if there is a moderator with k levels, this technique considers k models. Increasing the number of variables, the number of links among them also increases and their nature differs, yielding greater complexity. As we discussed for moderators, we also need to reflect on the nature of controls. Considering the case of only one control, if it is continuous, we simply introduce a new latent variable with the control as a unique indicator. However, this single-item approach is not free of criticism (Diamantopoulos et al, 2012). When the control is categorical, we can adopt the MGA using the levels to generate groups. Then, the overall model is compared in subgroups, and all the relationships are involved in the comparison, making it impossible to set the control only on a single link. The use of the MGA relies on the sample size; also when using PLS-SEM, which is applicable to a small sample size, a fitted issue could occur when we split in many subgroups, and it increases with a moderator. Considering Eq.2, the previously possible proposed solutions refer only to the single-item approach, which examines only one link at a time. In contrast, applying the MGA disables the choice of a solution since the MGA offers a complete picture of the control's influence on the system, considering the impact of the control on all the links. When theoretical assumptions require assessment with several controls, more than one approach is available. We can deal with controls separately, in this way we go back to previous cases, or conjointly. In the latter approach, we can alternatively apply the MGA, simultaneous single-item variables or one latent variable approach (also called a multi-item approach). With the MGA, we compare several subgroups identified by different mixtures of levels, considering the worsened issue of sub-sample dimensions discussed above. With the simultaneous single-item approach, we input as many latent variables as controls in the same model, with consideration to the worsened issue of single-item measure. Using single-item variables simultaneously or separately, as above, implies the same differences as in regression models when considering a unique multiple regression with all the controls together or several regressions that differ only by the control. Finally, we can identify a new construct using one latent variable measured by several controls. This approach requires accuracy and a choice between a reflective and formative shape to define the new construct. Either way, it is not easy to imagine a latent variable which summarizes different controls. To better explain the different options described, we introduce an illustrative example with *customer satisfaction* modelled as a mediator between *corporate reputation* and *customer loyalty* (see Eq.3). Supposing to control for *gender* and *age*, two research strategies are possible. In the first, the controls are separately considered. Treating *age* as continuous, a new latent variable (with single-item approach) must be added in the model, only on the dependent variable *loyalty* or on the *satisfaction* too. The same holds for *gender*, if it is handled as dichotomized. Considering *age* or *gender* as categorical, MGA has to be applied; in this case the overall model is controlled. A second research strategy analyses the controls simultaneously. With both the controls categorical, MGA implies to split the dataset in a number of sub-samples equal to the product

of their levels (e.g. for three *age* levels and two *gender* levels, six sub-samples are considered). Taking into account *age* as continuous and *gender* as categorical, two sub-samples are created, for female and male, and with *age* treated in the way described above. If *age* and *gender* are expressed as continuous or dichotomized, two different latent variables can be inserted simultaneously in the same model, both of them or alternatively only on *loyalty* or on *satisfaction* too. Finally, the multi-item approach, with an unique latent variable summarizing all the controls, in the specific case of *gender* and *age* is not suitable.

3 Discussion

Guidelines to consider control variables in PLS-SEM could contribute to the literature debate. First of all, it is essential to pay attention to the recommendations of Becker et al (2016) about controls selection: to provide a brief explanation for why each control was selected, to avoid proxies for them, to include them in hypotheses and to describe why and how each covariate was measured in the analysis, noting that a clear description allows for effective replication and extension of the findings. Since in PLS-SEM preliminary analysis we usually evaluate the constructs by applying exploratory factor analysis, it could be helpful to test the correlations between the first factors of the considered constructs with the hypothesised theoretical controls. In this way, an initial choice could be made. After this first step, we can input controls into the model in different ways. In PLS-SEM we have variables that can be simultaneously dependent and independent. Therefore, there is the issue of choosing the variables to which the controls must be related. If we use the MGA, this problem does not arise since we are analysing the model as a whole. This approach allows us to simultaneously consider more than one control, but it could be unrealistic because of the sample size; moreover, the results could be difficult to interpret. As an alternative to the MGA, we can adopt the single- or multi-item approach. With many controls, the single-item approach can be implemented, examining the controls one by one or simultaneously. If several controls are summarized in a latent variable, we have to check the reliability and validity of the construct. In all of these cases, we have to initially choose the relationships between the controls and the other variables; for this purpose, we need to take the theory into account, and we can conduct a comparative test to understand the role of the controls. Their effects can be evaluated by the intensity and significance of the estimated paths. The criterion of parsimony always applies to avoid inserting too many controls, which makes interpretation difficult. Referring to the advice of a report on the shared variance between original and residual predictor variables (Breaugh, 2008), while recognizing its remarkable benefits, we highlight that in PLS-SEM because of their complexity it is difficult to provide overall suggestions. This is still an open issue for future development. After testing the results with and without controls, if the results do not differ, only the analysis without controls should be displayed, along with a motivation of the removal. Moreover, statistics of the controls must be included in the

model results. Finally, researchers must be cautious with result generalizations involving controls. Indeed, by using them, a statistical sample of individuals is created for which predictors that are correlated are forced to be orthogonal or independent. Therefore, it is questionable to make inferences on the population.

References

- Baron RM, Kenny DA (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J PERS SOC PSYCHOL* 51(6):1173–1182
- Becker T (2005) Potential problems in the statistical control of variables in organizational research: A qualitative analysis with recommendations. *ORGAN RES METHODS* 8(3):274–289
- Becker T, Atinc G, Breugh J, Carlson K, Edwards J, Spector P (2016) Statistical control in correlational studies: 10 essential recommendations for organizational researchers. *J ORGAN BEHAV* 37(2):157–167
- Breugh J (2008) Important considerations in using statistical procedures to control for nuisance variables in non-experimental studies. *HUM RESOUR MANAGE R* 18(4):282–293
- Chin W, Marcolin B, Newsted P (2003) A partial least squares latent variable modeling approach for measuring interaction effects: Results from a monte carlo simulation study and an electronic-mail emotion/adoption study. *INFORM SYST RES* 14(2):189–217
- Diamantopoulos A, Sarstedt M, Fuchs C, Wilczynski P, Kaiser S (2012) Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. *J ACAD MARKET SCI* 40(3):434–449
- Hair J, Sarstedt M, Ringle C, Mena J (2011) An assessment of the use of partial least squares structural equation modeling in marketing research. *J ACAD MARKET SCI* 40(3):414–433
- Henseler J, Ringle C, Sinkovics R (2009) The use of partial least squares path modeling in international marketing. In: Rudolf R Sinkovics PNG (ed) *New Challenges to International Marketing*, vol 20, Emerald Group Pub. Lim., pp 277–319
- Judd C, Kenny D (1981) Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review* 5(5):602–619
- Lohmöller J (1989) *Latent Variable Path Modeling with Partial Least Squares*. Physica-Verlag Heidelberg
- Spector P, Brannick M (2011) Methodological urban legends: The misuse of statistical control variables. *ORGAN RES METHODS* 14(2):287–305
- Spector P, Zapf D, Chen P, Frese M (2000) Why negative affectivity should not be controlled in job stress research: don't throw out the baby with the bath water. *J ORGAN BEHAV* 21(1):79–95
- Wold H (1985) Partial least squares. In: Kotz S, Johnson N (eds) *Encyclopedia of Statistical Sciences*, John Wiley, New York, pp 581–591

Partial dependence with copula and financial applications

Dipendenza parziale con funzioni copula e applicazioni finanziarie

Giovanni De Luca, Marta Nai Ruscone and Giorgia Riveccio

Abstract In this paper the partial dependence with copula function is presented. In some context, the dependence structure between two variables can be highly influenced by one or more covariates, so it is of interest to know how this dependence structure changes with the value taken by the covariates. An application is carried out to study the dependence structure among eight financial assets. After analyzing the marginal dependence structure, we condition on the market index and finally both on the level and on the volatility of the market.

Abstract *Il focus di questo lavoro è la dipendenza parziale stimata con funzioni copula. In alcuni contesti, la struttura di dipendenza tra due variabili può essere fortemente influenzata da una o più covariate, quindi è interessante sapere come questa struttura di dipendenza cambia con il valore assunto dalle covariate. Un'applicazione a dati finanziari presentata con l'obiettivo di verificare il cambiamento della struttura di dipendenza tra otto attività finanziarie. Dopo aver analizzato la struttura di dipendenza marginale, si condiziona la stessa all'indice di mercato e, infine, sia al livello che alla volatilità del mercato.*

Key words: Copula function, conditioning, financial returns.

Giovanni De Luca

Università di Napoli Parthenope, via G. Parisi, 13 - 80132 Napoli
e-mail: giovanni.deluca@uniparthenope.it

Marta Nai Ruscone

LIUC - Università Cattaneo, Corso Matteotti 22 - 21053 Castellanza (VA)
e-mail: mnairuscone@liuc.it

Giorgia Riveccio

Università di Napoli Parthenope, via G. Parisi, 13 - 80132 Napoli
e-mail: giorgia.riveccio@uniparthenope.it

1 Copula functions

An effective way used to capture the dependence structure of a multivariate distribution is the copula distribution function. Although copulas can be defined for any multivariate distributions in \mathbb{R}^d , we focus on bivariate continuous random vectors for expository purpose.

Let denote $F_{Y_1, Y_2}(y_1, y_2)$ the bivariate cumulative distribution of the pair (y_1, y_2) of the random variables Y_1 and Y_2 , $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ the marginal c.d.f. of the Y_1 and Y_2 , respectively. As show in [4], the joint c.d.f. of (Y_1, Y_2) can be written as:

$$F_{Y_1, Y_2}(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2) = C(F_{Y_1}(y_1), F_{Y_2}(y_2))$$

where C is the c.d.f. of the distribution on $[0, 1]^2$, with the uniform margins. When variables are continuous, C is unique, and is called the copula of (Y_1, Y_2) . Sklar's theorem [4] allows to separate the marginal feature and the dependence structure which is represented by the copula. The function C is the c.d.f. of the pair (U, V) where $U = F_{Y_1}(Y_1)$ and $V = F_{Y_2}(Y_2)$, and

$$c(u, v) = \frac{\partial^2 C}{\partial u \partial v}(u, v)$$

is the associated p.d.f. Sklar's theorem proves the existence and the uniqueness of the copula. It also explains how to construct it from the initial distribution. Indeed, for any $0 \leq u, v \leq 1$, the copula is given by

$$C(u, v) = F_{Y_1, Y_2}(F_{Y_1}^{-1}(u), F_{Y_2}^{-1}(v)),$$

where $F_{Y_1}^{-1}$ and $F_{Y_2}^{-1}$ are the marginal quantile functions. The copula characterizes any nonlinear dependence which is invariant by increasing transformation of either Y_1 and Y_2 . More precisely we have the following: if ϕ and ψ are strictly increasing functions, then (Y_1, Y_2) and $(\phi(Y_1), \psi(Y_2))$ have the same copula.

However, the dependence structure between two variables can be highly influenced by some covariates, and it can be of interest to know how this dependence structure changes with the value taken by the covariates. This motivates the need for introducing conditional copulas, and the associated Kendall's τ measure of association. Conditional copulas have been formally introduced by [2, 3]. They are rather straightforward extensions of the latter concepts, when dealing with conditional distributions. Suppose that the conditional distribution of (Y_1, Y_2) given the values of the covariates fixed at a given level, say \mathbf{X} , exist and denote the corresponding conditional joint distribution function by

$$H_X(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2 | X = x).$$

If the marginals of H_X denoted as

$$F_{1X}(y_1) = P(Y_1 \leq y_1 | X = x)$$

and

$$F_{2X}(y_2) = P(Y_2 \leq y_2 | X = x)$$

are continuous, then according to Sklar's theorem [1] there exists a unique copula C_X which equals

$$C_X(u_1, u_2) = H_X(F_{1X}^{-1}(u_1), F_{2X}^{-1}(u_2)),$$

where $F_{1X}^{-1}(u) = \inf\{y : F_{1X}(y) \geq u\}$ is the conditional quantile function of Y_1 given $X = x$ and F_{2X}^{-1} is the conditional quantile function of Y_2 given $X = x$. The conditional copula C_X fully describes the dependence structure of (Y_1, Y_2) given $X = x$.

The associated Kendall's τ measure is given by:

$$\tau_X(u_1, u_2) = \tau(u_1 | X, u_2 | X).$$

2 Analysis of financial returns

The marginal and partial dependence structure of a set of n assets has been analyzed. In particular, we have considered $n = 8$ assets traded in the Italian stock exchange (A2A, BPER, Enel, FCA, Generali, Intesa San Paolo, STM, Unicredit) for the period February 27th 2013 - January 30th, 2018. Given $n = 8$, we have $n(n-1)/2 = 28$ pairs of assets. We have first selected the best copula that describe each bivariate distribution among Independence copula, Gaussian copula, Student's t copula, BB1 copula and BB7 copula. According to the tail dependence, Gaussian copula does not admit tail dependence, Student's t copula allows for symmetric tail dependence, while for BB1 and BB7 copula the tail dependences can be different. The selection of the copula function has been based on the AIC criterion.

The dependence structure in terms of Kendall's τ is showed in Table 1. All the values are positive, as the stylized facts of financial returns suggest.

Table 1 Marginal dependence structure.

	BPER	Enel	FCA	Generali	ISP	STM	Unicredit
A2A	0.377	0.313	0.447	0.162	0.173	0.304	0.316
BPER		0.430	0.464	0.196	0.190	0.293	0.279
Enel			0.337	0.147	0.157	0.228	0.211
FCA				0.213	0.220	0.370	0.356
Generali					0.307	0.294	0.285
ISP						0.379	0.347
STM							0.604

The second step is based on the study of the dependence structure of the assets conditionally on the level of the market index FTSEMIB (Table 2). It is evident that there is a strong weakening of the dependence structure. Moreover, the highest unconditional Kendall's τ has been observed for the pair STM/Unicredit while, when

the conditioning on the level of the FTSEMIB index is taken into account, the pair BPER/Enel shows the highest value.

Table 2 Partial dependence structure. The conditioning variable is the level of the FTSEMIB index.

	BPER	Enel	FCA	Generali	ISP	STM	Unicredit
A2A	0.201	0.200	0.267	−0.015	−0.051	0.051	0.053
BPER		0.312	0.253	0.000	−0.065	−0.056	−0.082
Enel			0.185	0.000	−0.022	−0.023	−0.043
FCA				−0.011	−0.051	0.015	−0.011
Generali					0.079	−0.095	−0.075
ISP						−0.074	−0.075
STM							0.190

Finally, the third step consists in the analysis of the dependence structure of the assets conditionally on the level of the FTSEMIB index as well as on its absolute value which is a measure of the volatility (Table 3). In this case the dependence structure does not show an appreciable change, which implies that the volatility of the market, measured as absolute value of the market index, is not effective in addition to the conditioning on the level of the market index.

Table 3 Partial dependence structure. The conditioning variables are the level of the FTSEMIB and its absolute value.

	BPER	Enel	FCA	Generali	ISP	STM	Unicredit
A2A	0.200	0.198	0.265	−0.015	−0.052	0.053	0.058
BPER		0.313	0.258	0.000	−0.061	−0.039	−0.063
Enel			0.182	0.000	−0.026	−0.024	−0.035
FCA				−0.014	−0.054	0.003	−0.009
Generali					0.071	−0.089	−0.073
ISP						−0.077	−0.075
STM							0.171

3 Conclusions

The dependence structure between two variables can be highly influenced by one or more covariates. Taking into account the conditioning variable can lead to a different perspective of the relationship among financial assets. As a result, further analyses, such as clustering and portfolio allocation, can be seriously affected.

References

1. Nelsen, R.: An introduction to copulas. 2nd ed. Springer, New York (2006)
2. Patton, A.: Modelling asymmetric exchange rate dependence. *Internat. Econom. Rev.* **47**, 527–556 (2006)
3. Patton, A.: Estimation of multivariate models for time series of possibly different lengths. *J. Appl. Econometrics* **21**, 147–173 (2006)
4. Sklar, A.: Fonctions de repartition a n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8**, 229–231 (1959)

Exploring the relationship between fertility and well-being: What is smart?

Esplorando la relazione tra fecondità e benessere: cosa c'è di smart?

Alessandra De Rose, Filomena Racioppi, Maria Rita Sebastiani

Abstract We analyse the association between fertility rates and well-being in the Italian regions in the period 2012-2017. Well-being is measured by the indicators of Equitable and Sustainable Well-being (BES), collected by ISTAT since 2013. We expect that the regions performing better in terms of well-being conditions are also those with the highest level of fertility. We use graphical and statistical methods to compare the rankings of Italian regions according to different BES domains with that according to TFR. The correlation analysis supports our main hypothesis; some targeted regression analyses let us to identify those well-being dimensions that matter more in explaining fertility variation.

Abstract Analizziamo l'associazione tra tassi di fecondità e benessere nelle regioni italiane nel periodo 2012-2017. Il benessere viene misurato attraverso gli indicatori di Benessere Equo e Sostenibile (BES), raccolti dall'Istat dal 2013. Ci aspettiamo che le regioni che si comportano meglio in termini di condizioni di benessere siano anche quelle con il più alto livello di fecondità. Utilizziamo metodi grafici e statistici per confrontare le graduatorie delle regioni italiane in base ai diversi domini BES e quella relativa al TFR. L'analisi delle associazioni supporta la nostra ipotesi principale; analisi di regressione mirate permettono di identificare quelle dimensioni di benessere che contano di più nello spiegare la variazione della fecondità.

Key words: Fertility, Well-being, BES

Introduction

In Italy, the period fertility level - measured by the mean number of children per woman or TFR – continues to decline (1.32 in 2017), by far below both replacement

¹ Alessandra De Rose, Sapienza University of Rome; email: alessandra.derose@uniroma1.it
Filomena Racioppi, Sapienza University of Rome; email: filomena.racioppi@uniroma1.it
Maria Rita Sebastiani, Sapienza University of Rome; email: mariarita.sebastiani@uniroma1.it

Exploring the relationship between fertility and well-being: what is smart?

level and desired number of children (ISTAT, 2014). Although the regional differences are very attenuated, there are pockets of great severity (very low fertility: 1.06 in Sardinia) but also situations of relative virtuosity (fertility in recovery: 1.74 in the Province of Bolzano). The same regions rank very distant by level of well-being as measured by the indicators of Equitable and Sustainable Well-being (BES in Italian; ISTAT, 2018). Thus, one cannot but wonder about the role played by the context in which individuals live on their life choices and about the effect of implicit and explicit policies that, improving the citizens quality of life, also favor the realization of their reproductive expectations. This, in turn, makes the level of fertility increase with a positive effect on population balance and structure.

The goal of this work is to study the association between well-being measured at aggregate level in terms of social, economic, environmental conditions and the reproductive behavior of the population measured by fertility. The main hypothesis is that the regions with the highest values in terms of welfare conditions are also those with the highest level of fertility.

2. Background and Hypotheses

Many studies evaluate the role of the contextual factors on population reproductive behaviour (see Entwisle, 1991 for a review). Meggiolaro (2011) in a study on Milan found that the characteristics of the place where people live influence their reproductive behaviours, even after controlling for individual factors. The changing social and economic conditions of a territory play a role in explaining differences in fertility. The recent economic downturn drew attention to the different reaction across European countries in terms of aggregate fertility (Goldstein et al., 2013). For Italy, Cazzola et al. (2016) found that, while in Northern and Central regions the rise of unemployment rate during the economic crisis related to a reduction in fertility levels, in the South the relationship has been weaker. Overall, the residential context helps explaining fertility diversity (Kulu and Washbrook, 2014). However, little is known about the mechanisms behind the geographical variability in reproductive outcomes (Fiori et al., 2013), as well as about the specific role played by the different dimensions of the regional well-being.

Well-being is a complex and multidimensional concept, that involves many social and environmental dimensions. It is linked to the available resources, the quality of life, subjective well-being, equity and sustainability. One single indicator – namely GDP – is not enough to represent it and it is necessary to assess a measurement system. For over 20 years, OCSE, European Commission, United Nations, World Bank and other international organizations have been promoting initiatives to sensitize governments to go beyond the economic production dimension. The rationale is to measure well-being by a multidimensional approach that includes, on one hand, subjective evaluations and perceptions, and, on the other hand, socio-economic conditions and environmental sustainability.

Italy is ahead in defining and estimating well-being indicators. In 2011, CNEL, the National Council for Economics and Labour, and ISTAT began a collaboration,

Exploring the relationship between fertility and well-being: what is smart?

involving many social actors (institutions, citizens, companies, labour unions). The result was the definition of a system of indicators of the quality of life, measuring Equitable and Sustainable Well-being (BES) and organized in 12 thematic domains at regional and provincial level. Since 2013, ISTAT annually publishes the BES Report; in 2018 the sixth edition.

Here we are interested in the association at regional level between the fertility behaviour and the regional context of well-being. We are also interested in changes through time in the relationship itself. Since the recent economic downturn produced negative effects both on fertility and social welfare, we analysed the associations between BES and TFR in two points in time, during and after the economic recession. Namely, we took the regional total fertility rate of the years 2012 and 2017 and related them to the average value of the BES composite indicators of the two years before, 2010-11 and 2015-16 respectively. We lagged the well-being indicators of two years in order to account for the time length which is needed for a certain contextual variation to produce an effect on fertility decisions.

3. Data

In the BES proposal, well-being is treated like a multidimensional concept and measured by means of a huge number of elementary indicators (around 130) grouped in 12 domains, each dealing with a domain of human well-being. It is a dynamic measurement system: every year the indicators are updated (confirmed or substituted) to account for any change in their trends or in the data source. The 12 domains pertain: health, education, work life and balance, economic well-being, subjective well-being, social relationships, safety, politics and institutions, landscapes and cultural heritage, environment, quality of services, innovations research and creativity. We propose the use of the effective syntheses of the BES indicators, that is the composite indicators built with the AMPI method (Mazziotta and Pareto, 2015). We selected the BES composite indicators at regional level as provided by ISTAT, for the period 2010-2016, and the TFRs for the years 2012-2017.

4. Methods

The geographical variability of the TFR and that of the BES composite indicators has been analysed for all the years since 2010 to 2016. In addition, we produced regional maps representing, for each composite indicators and for TFR, the mean values calculated over the period 2010-2016. Afterwards, we applied descriptive methods to measure the degree of similarity between the fertility map and each of the specific-domains BES maps. Specifically, to compare the rankings of Italian regions obtained according to the different BES domains with the ranking according to the TFR, we applied the Spearman correlation coefficient.

Exploring the relationship between fertility and well-being: what is smart?

Although the results of the analysis referred to the overall period confirm our main hypothesis (see Section 5), we noticed that both TFR and BES indicators show significant variations through time. Thus, we suppose that the associations between fertility and each of the well-being indicators could vary over time by sign and/or by intensity. We are specifically interested in comparing the situation during and after the Great Recession. For this reason, we deepened the study by applying the statistical analyses twice corresponding to two different years, and then comparing the obtained results. Namely, we focussed on the fertility levels in 2017 and on that registered five years before in 2012 and put them in relation with the average BES composite indicators values calculated over the years 2015-2016 and 2010-2011, respectively. We performed the most adequate correlation tests. Finally, we applied some targeted regression analyses in order to identify the BES domains that show a significant effect on the TFR.

5. Results

Figure 1 shows the mean values of the BES composite indicators for the whole period 2010-2016 in the three main Italian geographical areas¹. Maps for a selection of BES composite indices and for TFR are reported in Figure 2: colours represent the distribution quartiles, with red representing the worst situation according to the specific domain and brilliant green the best. The results support our main hypothesis: the regions where all the well-being indicators are at the highest value (i.e. Trentino Alto Adige) shows the highest level of fertility, while those with the lowest fertility level (and Sardinia among them), perform worse as far as many BES indicators is concerned. The regional maps of economic well-being, employment and quality of services are those more overlapping that of fertility according to the analysis of similarity between each composite index and TFR.

Finally, table 1 shows the results of the two separated correlation analyses as described in section 4². In both periods, the correlations between almost all the BES composite indicators and fertility at regional level are positive and significant. The fertility in 2012 was mainly associated with the economic and employment situation of the two years before. Instead, fertility in 2017 has been mainly associated with subjective conditions. The different role played by the indicators in the two periods is confirmed by a preliminary OLS: during the crisis, regional fertility levels appeared to have been mainly influenced by objective economic conditions, while, right after the hardest times, subjective well-being has been the most important predictor of fertility.

¹ Note that all the composite indices have the same positive polarity with respect to well-being and that, for the sake of comparison in space and time, Italy 2010=100 for each composite index.

² Before performing the correlation analysis, we tested the normality assumption by means of the Shapiro Wilks test.

6. Discussion

Our results evoke the existence of a non-negligible relationship between the well-being of the territory and the reproductive behavior of the population. The message for the policy is quite straightforward: increasing people's quality of life enables them to carry out their projects in terms of family building. This can be a *smart* choice at local level. We also suggest that the use of effective syntheses of the BES indicators and performing a system of simple, though powerful methods of analysis help building a *smart* framework for the study of the relationship between fertility and well-being at territorial level.

Further steps of this study will include a more refined regression analyses in order to identify more informative indicators for fertility variability, also recognizing the spatial nature of the data. Furthermore, we will exploit the possibility to replicate the analyses using data at the provincial level.

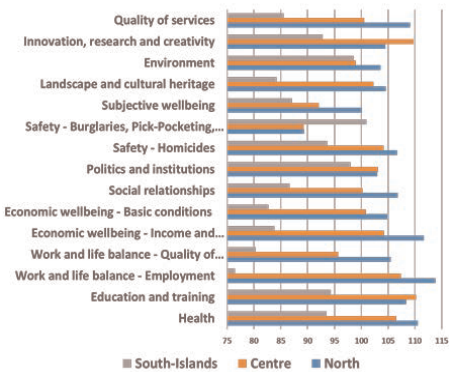


Figure 1: BES composite indicators by geographical areas. Average 2010-2016.

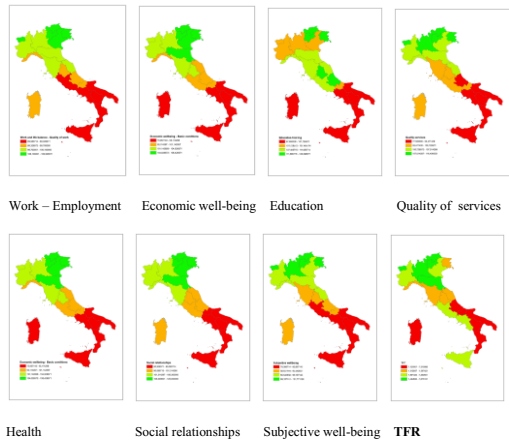


Figure 2: Maps of selected BES composite indicators and TFR. Average 2010-2016

Exploring the relationship between fertility and well-being: what is smart?

Table 1: Correlation analysis – TFT (2012 and 2017) versus BES composite indicators

BES Composite indicator	correlation between BES 2010-2011 (mean values) and TFT 2012			correlation between BES 2015-2016 (mean values) and TFT 2017		
	sample estimates	test		sample estimates	test	
Health	0,570	t = 3.0204**	(p-value = 0.007036)	0,550	t = 2.8738**	(p-value = 0.009722)
Education and training	0,301	t = 1.376	(p-value = 0.1848)	0,407	t = 1.9432	(p-value = 0.06696)
Work and life balance - Employment	0,671	t = 3.9419***	(p-value = 0.0008749)	0,487	t = 2.4279*	(p-value = 0.02529)
Work and life balance - Quality of work	0,566	t = 2.9959**	(p-value = 0.007429)	0,434	t = 2.1004*	(p-value = 0.04928)
Economic wellbeing - Income and Inequalities	0,698	t = 4.2504***	(p-value = 0.0004326)	0,517	t = 2.6347*	(p-value = 0.01633)
Economic wellbeing - Basic conditions	0,544	t = 2.8295*	(p-value = 0.01071)	0,341	t = 1.581	(p-value = 0.1304)
Social relationships	0,625	t = 3.4918**	(p-value = 0.00244)	0,627	t = 3.5059**	(p-value = 0.002363)
Politics and institutions	0,626	t = 3.4988**	(p-value = 0.002402)	0,590	t = 3.1859**	(p-value = 0.004866)
Safety - Homicides	$\rho = 0.6364855$ (Spearman test) S = 559.8124**			0,144	t = 0.6338	(p-value = 0.5338)
Safety - Burglaries, Pick-Pocketing, Robberies	-0,237	t = -1.0656	(p-value = 0.3)	-0,149	t = -0.6548	(p-value = 0.5204)
Subjective wellbeing	0,592	t = 3.2002**	(p-value = 0.004712)	0,736	t = 4.7337***	(p-value = 0.0001445)
Landscape and cultural heritage	0,621	t = 3.4537**	(p-value = 0.002661)	0,575	t = 3.0644**	(p-value = 0.006382)
Environment	0,470	t = 2.3227*	(p-value = 0.03145)	0,536	t = 2.7708*	(p-value = 0.01217)
Innovation, research and creativity	0,400	t = 1.911	(p-value = 0.07257)	0,190	t = 0.8428	(p-value = 0.4098)
Quality of services	0,630	t = 3.5371**	(p-value = 0.002202)	0,509	t = 2.5795*	(p-value = 0.01837)

Note: * = coefficient significant at p-value<=0.010; ** = coefficient significant at p-value<=0.005; *** = coefficient significant at p-value<=0.001

References

1. Cazzola A., Pasquini L., Angeli A.: The relationship between unemployment and fertility in Italy: A time-series analysis. *Demographic Research*, 34, 1-38 (2016)
2. Coale A.J., Watkins S.C. (eds): *The decline of fertility in Europe*. Princeton University Press, Princeton (1986)
3. De Rose A., Racioppi F., Zanatta A.: Italy: Delayed adaptation of social institutions to changes in family behavior. *Demographic Research*, 19, 665-704 (2008)
4. Entwisle B.: Micro-macro theoretical linkages in social Demography: A commentary, in Huber J. (ed), *Macro-micro linkages in Sociology*, pp. 280-286. Sage Publications, Newbury Park (1991)
5. Fiori F., Graham E., Feng Z.: Geographical variations in fertility and transition to second and third birth in Britain. *Advanced in Life Course Research*, 21, 149-167 (2013)
6. Goldstein J. R., Kreyenfeld M., Jasilioniene A., Örsal D.D.K.: Fertility reactions to the ‘Great Recession’ in Europe. *Demographic Research*, 29(4), 85-104 (2013)
7. ISTAT: *Avere figli in Italia negli anni 2000*, ISTAT, Roma (2014)
8. ISTAT: *Rapporto BES 2018 – Il benessere equo e sostenibile in Italia*. ISTAT, Roma (2018)
9. Kulu H., Washbrook E.: Residential context, migration and fertility in a modern urban society. *Advances in Life Course Research*, 21, 168-18. (2014)
10. Mazziotta M., Pareto A.: On a generalized non-compensatory composite index for measuring socio-economic phenomena. *Social Indicators Research*, 127(3), 983-1003 (2015)
11. Meggiolaro S.: Do neighborhoods have an influence on reproductive intentions? Empirical evidence from Milan. *Regional Studies*, 45(6), 791-807 (2011)

Web-Based Data Collection and Quality Issues in Co-Authorship Network Analysis

Qualità dei dati bibliografici raccolti via web per l'analisi di reti di collaborazione scientifica

Domenico De Stefano, Vittorio Fuccella, Susanna Zaccarin

Abstract In this contribution we discuss data quality issues related to the application of web scraping techniques to the Cineca IRIS platform to derive co-authorship data among Italian university scholars. First, a semi-automatic tool is adopted to retrieve metadata from the platform, then a disambiguation network-based approach is considered to deal with author name disambiguation. This combined procedure is used to derive the co-authorship relations among Italian academic statisticians on the basis of the publications they inserted in the IRIS system until 2017.

Abstract *Il presente contributo discute aspetti della qualità dei dati su pubblicazioni scientifiche ottenuti dall'interrogazione con tecniche di web scraping della piattaforma IRIS del Cineca al fine di costruire reti di collaborazione scientifica. Alcuni risultati relativi alla rete di collaborazione degli statistici universitari italiani sono riportati.*

Key words: co-authorship network, IRIS platform, web scraping, disambiguation algorithm

1 Introduction

Scientific collaboration is an important driver of research progress that supports researchers in the generation of novel ideas. It has been also recognized as a key factor in measuring and evaluating scientific performance of scholars. Among the

Domenico De Stefano

Department of Political and Social Sciences, University of Trieste e-mail: ddestefano@units.it

Vittorio Fuccella

Department of Informatics, University of Salerno e-mail: vfuccella@unisa.it

Susanna Zaccarin

Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste e-mail: susanna.zaccarin@deams.units.it

widespread applications of Social Network Analysis (SNA) in the last decades, the study of co-authorship networks, used as a proxy of scholars collaborative behavior, is one of the topic that most benefited from SNA perspective. Seminal studies explored co-authorship networks in various fields using data gathered from large online international Digital Libraries (DLs) - general (e.g., ISI-WOS, Scopus) or thematic oriented (e.g., Econlit for Economics or Medline for Medical Sciences) - rather than collected by interviews or questionnaires administered directly to the authors of the papers. Another stream of research focuses on interactions among members of a given target population (e.g., scholars involved in a scientific community or affiliated to a given institution) in order to retrieve the pattern of collaborative behaviors and its effect on the scholars scientific performance. In this case, recent literature pointed out that international DLs provide a partial coverage of the entire scholar scientific production as well as under coverage of a target population (6; 7; 5), since writing articles or books, publishing in international or national oriented journals, can be highly dependent on discipline specialty (4).

In this framework, scientific production of the Italian academic scholars can be retrieved from individual web pages (“sito docente”), managed by the Italian Ministry of University and Research (MIUR) and the Cineca consortium. Unfortunately, the access to this database is not freely available, due to privacy policies.

The recent introduction of the Institutional Research Information System (IRIS) developed by Cineca consortium seems to furnish a unique platform in Italy for managing and supporting research in academic and research institutions. Within this system, it is available an open archive module for the repository of the research products allowing the storage, the consultation and the enhancement of these outputs. Thanks to this tool, the affiliated universities can access to a system able to communicate with the national (i.e. “sito docente” of MIUR) and international databases for the management and dissemination of scholars’ scientific publications. The platform presents both pros and cons common to other national-based DLs. Even if it guarantees a high coverage rate of our target population and its scientific production, to retrieve co-authorship ties among scholars it is necessary to combine the data contained in different platform deployments available at each university. In addition, data quality is affected by the manual publication data entry made by authors. Moreover, no details are provided on co-authors external to the target population, which implies a huge effort in author name disambiguation. To deal with these aspects, we first propose a web scraping procedure based on a semi-automatic tool retrieving publication metadata from the online platform in order to reduce the manual adjustments. Second, we introduce a network-based approach to deal with author name disambiguation that requires a minimal set of record attributes (identifier, co-authors, venue) to obtain clean data set from which we can construct a proper co-authorship network. We apply this combined procedure to derive co-authorship relationships among Italian academic statisticians, considering the publications they inserted in the IRIS system until the 2017.

The paper is organized as follows. Section 2 reports details on the web scraping procedure adopted to extract data from the IRIS system. First results obtained from

the name disambiguation procedure are discussed. Section 3 contains some final remarks.

2 Web scraping techniques to extract publications of Italian statisticians

We aim at reproducing the co-authorship network among the Italian academic statisticians that is, those scientists classified as belonging to one of the five subfields established by the governmental official classification: Statistics (Stat), Statistics for Experimental and Technological Research (Stat for E&T), Economic Statistics (Economic Stat), Demography (Demo), and Social Statistics (Social Stat). This target population is composed of the 721 statisticians, as recorded in the MIUR database at July 2017 (2).

Since most of the Italian universities adopt the IRIS platform for publications data storage (70 out of 97 total institutions, <https://www.cineca.it/en/content/IRIS-institutional-researchinformation-system>), we use a web-scraping data techniques based on a semi-automated tool to retrieve the whole publication metadata of our target population. In principle, each author has a page from which it is possible to access to the data of his/her publications. The semi-automated tool uses as input a table containing references (name, surname and academic institution) of the 721 statisticians. Moreover, the tool is programmed –in Java– with the aim of automatically extract the data from the system obtaining a good coverage of the author publications and reducing the manual adjustments to manage errors or uncertainty conditions.

It is worth noting that each institution hosts a different deployment of the system, thus each statistician is linked to the index page of the IRIS deployment of his/her institution. The author's page contains the list of publications of which the person is co-author. If an author has more than 100 publications, these are necessarily split into multiple pages. Each publication in the list is associated to a link to a new page containing the details of the publication (title, authors, venue, year of publication and various identifiers –URL, DOI, ISI codes WoS and Scopus and so on). The proposed web scraping procedure retrieved and followed the links of each author publication in order to download these metadata. The procedure allowed us to obtain a coverage around 80% of the target population of statisticians. This rate is in line with those obtained in previous studies (1). The complete database of publication records for each author could contain many duplications in presence of publications co-authored by statisticians affiliated to different institutions. Indeed, in each IRIS system the same publication can be reported with a different format. Before the co-authorship network construction we addressed both product and author name duplications by adapting to this context the procedure proposed in (3).

As an example of the author disambiguation algorithm we focus on the subset of Social Statisticians field for which we retrieve 4258 publications and 3636 total

number of authors of which 778 are internal to the IRIS system (i.e., they are Italian academic scholars affiliated to various scientific fields).

The goal here is to consolidate the identities spelled in different ways because of the authors manual input of IRIS metadata for all the 3636 authors.

The disambiguation procedure use a graph-based representation of author occurrences, each of which is associated to a graph vertex. We added an edge between the two vertices every time their associated occurrences showed some evidence of belonging to the same identity. The output identities were obtained by calculating the connected components of the graph, each connected component being a different identity.

An edge was added between two vertices if their associated occurrences were compatible and showed at least one of the following evidences, based on the attributes of their respective publication records:

- at least one co-author in common;
- same publication venue;
- the titles shared at least one keyword.

For each checked occurrence, the associated vertex is only connected to the vertex with the highest evidence. We calculated an evidence measure as:

$$E = w_a \times e_a + w_v \times e_v + w_t \times e_t$$

where E has real values in the range $[0, 1]$; w_a , w_v , and w_t are the weights for the functions e_a , e_v , and e_t , respectively.

Fig 1 shows the results obtained for two identities related to two authors, namely ‘Domenico De Stefano’ and ‘Maria Prosperina Vitale’. For the first identity the procedure works well, correctly merging the occurrences. However the procedure fails when the name is severely misspelled and there is a lack of evidence related to both common co-authors and publication keywords, likewise in the case of ‘Vitale M’ for the identity ‘Maria Prosperina Vitale’.

3 Final remarks and future lines of research

The availability of the IRIS archive is certainly a promising tool but a lot of issues must be managed during the data collection process from this source. The unavailability of IRIS platform in some universities, the private access to the IRIS system in some cases, the different publication data format, and the presence of more than one record found for the same author are examples of errors obtained after the extraction of information. In addition, author disambiguation processes needs to be taken into account to obtain co-authorship data among scholars affiliated in different universities.

After the data cleaning to reconcile publication records, the detection of duplicates and the recognition of internal and external authors of the same publications, the co-authorship networks will be derived.

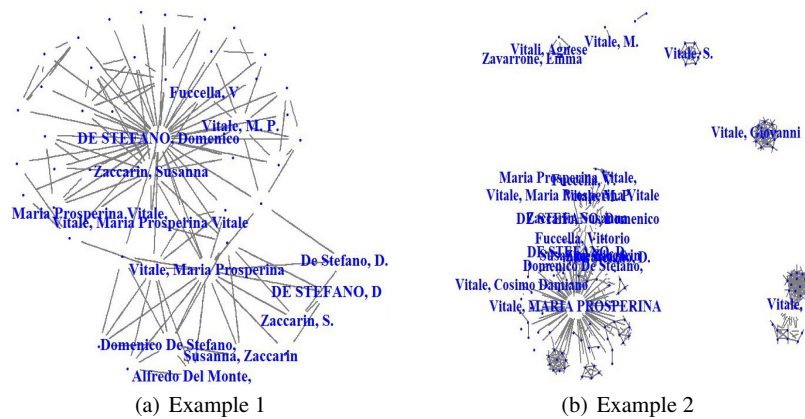


Fig. 1 Results of the author name disambiguation procedure for two author identities, ‘Domenico De Stefano’ (example 1) and ‘Maria Prosperina Vitale’ (example 2).

References

- [1] De Stefano, D., Fuccella, V., Vitale, M. P., Zaccarin, S.: The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*. **35**, 370-381 (2013)
- [2] De Stefano, D., Fuccella, V., Vitale, M. P., Zaccarin, S.: Using web scraping techniques to derive co-authorship data: insights from a case study. In: Abbruzzo, A., Brentari, E., Chiodi, M., Piacentino, D. (eds.) *Book of short Papers SIS 2018*, Pearson (2018)
- [3] Fuccella, V., De Stefano, D., Vitale, M. P., Zaccarin, S.: Improving co-authorship network structures by combining multiple data sources: evidence from Italian academic statisticians. *Scientometrics*. **107**, 167-184 (2016)
- [4] Hicks, D.: The difficulty of achieving full coverage of International Social Science literature and the bibliometric consequences. *Scientometrics* 44, 193-215 (1999)
- [5] Mitchell, R.: *Web scraping with Python: collecting data from the modern web*. Packt Publishing, Birmingham (2015)
- [6] Murthy, D., Gross, A., Takata, A., Bond, S.: Evaluation and development of data mining tools for social network analysis. In: Ozyer, T., Erdem, Z., Rokne, J., Khoury, S. (eds.) *Mining Social Networks and Security Informatics*, pp. 183-202. Springer, Dordrecht (2013)
- [7] Vargiu, E., Urru, M.: Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Artificial Intelligence Research*. **2**, 44-54 (2013)

A new regression model for bounded multivariate responses

Un nuovo modello di regressione per risposte multivariate limitate

Agnese Maria Di Brisco, Roberto Ascari, Sonia Migliorati and Andrea Ongaro

Abstract The aim of this work is to propose a new multivariate regression model for compositional data, i.e., vectors of proportions. It is based on a mixture of Dirichlet-distributed components and it enables many relevant properties for compositional data as well as accounting for positive correlations. Despite the complexity of the model, its special mixture structure provides a greater flexibility and a richer parameterization than the standard Dirichlet regression (DirReg) model and, moreover, guarantees its identifiability. We illustrate the performance and the goodness of fit of our new model through an application to the last Italian elections data.

Abstract *L'obiettivo di questo lavoro è quello di proporre una nuova regressione multivariata per dati composizionali, ossia vettori di proporzioni. Essa è basata su una mistura di componenti aventi distribuzione Dirichlet ed è caratterizzata da numerose proprietà rilevanti per dati composizionali, oltre ad ammettere correlazioni positive. Nonostante la complessità del modello, la sua speciale struttura a mistura garantisce una maggiore flessibilità ed una più ricca parametrizzazione rispetto alla regressione Dirichlet e, inoltre, ne garantisce l'identificabilità. Un'applicazione ai dati delle ultime elezioni italiane consente di mettere in luce l'ottima bontà di adattamento del nuovo modello.*

Key words: simplex, mixture model, dirichlet distribution, bayesian inference.

Di Brisco Agnese Maria
University of Milano-Bicocca, e-mail: agnese.dibrisco@unimib.it

Ascari Roberto
University of Milano-Bicocca e-mail: roberto.ascari@unimib.it

Migliorati Sonia
University of Milano-Bicocca e-mail: sonia.migliorati@unimib.it

Ongaro Andrea
University of Milano-Bicocca e-mail: andrea.ongaro@unimib.it

1 Introduction

Compositional data, namely proportions of some whole, are defined on the simplex space $\mathcal{S}^D = \{\mathbf{Y} : Y_j > 0, j = 1, \dots, D, \sum_{j=1}^D Y_j = 1\}$. A proper distribution for a D -dimensional vector on \mathcal{S}^D is the Dirichlet, $\mathbf{Y} \sim \text{Dir}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D)^\top$ and $\alpha_j > 0$. Since the aim of this work is to evaluate some regression models for compositional data, it is worth introducing an alternative parameterization based on mean-precision parameters, i.e.:

$$\begin{cases} \bar{\alpha}_j = \mathbb{E}(Y_j) = \frac{\alpha_j}{\alpha^+}, & j = 1, \dots, D \\ \alpha^+ = \sum_{j=1}^D \alpha_j \end{cases} \quad (1)$$

with $0 < \bar{\alpha}_j < 1$, $\sum_{j=1}^D \bar{\alpha}_j = 1$ and $\alpha^+ > 0$.

Despite the Dirichlet distribution has some statistical properties, it is affected by some limitations [1]. By way of example, it imposes a very rigid structure on the covariance matrix, and it fails to model several scenarios among which multimodal shapes of the distribution. To overcome these drawbacks, we take advantage of a more flexible distribution for compositional data, i.e., the Extended Flexible Dirichlet (EFD) [9, 11], and we propose a new regression model based on it.

The EFD distribution is a special finite mixture with Dirichlet components. The distribution function of an EFD distributed random vector \mathbf{Y} admits the following mixture representation:

$$\text{EFD}(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \sum_{r=1}^D p_r \text{Dir}(\mathbf{y}; \boldsymbol{\alpha} + \tau_r \mathbf{e}_r), \quad (2)$$

where $\text{Dir}(\cdot; \cdot)$ denotes the Dirichlet distribution, \mathbf{y} and \mathbf{p} lie in \mathcal{S}^D , $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D)^\top$, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_D)^\top$, $\alpha_r > 0$, $\tau_r > 0$, and \mathbf{e}_r is a vector of zeros except for the r -th element which is equal to one. Its probability density function (p.d.f.) can be written as:

$$f_{\text{EFD}}(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \left(\prod_{r=1}^D \frac{y_r^{\alpha_r-1}}{\Gamma(\alpha_r)} \right) \sum_{h=1}^D p_h \frac{\Gamma(\alpha_h) \Gamma(\alpha^+ + \tau_h)}{\Gamma(\alpha_h + \tau_h)} y_h^{\tau_h}. \quad (3)$$

The EFD distribution contains the Dirichlet as an inner point when $\tau_r = 1$ and $p_r = \bar{\alpha}_r$ for every $r = 1, \dots, D$ (see (1)). Thanks to its mixture structure, the EFD allows for a large variety of shapes of its p.d.f., including uni- and multi-modal ones. Furthermore, due to its rich parameterization, the EFD allows for a flexible modelization of the dependence structure of the composition, overcoming the drawbacks imposed by the Dirichlet distribution, and making it an interesting alternative to classical distributions for compositional data. Last, the EFD distribution has several theoretical properties among which some simplicial forms of dependence/independence [9, 11] and identifiability [3].

2 A new regression model based on a reparameterized EFD

Since our aim is to propose a new regression model based on the EFD, we first have to introduce a new parameterization that explicitly includes the mean vector. To this end, according to the mean-precision parameterization in (1), we observe that the r -th Dirichlet component in (2) has a mean vector $\boldsymbol{\lambda}_r$ equal to:

$$\boldsymbol{\lambda}_r = (1 - w_r)\bar{\boldsymbol{\alpha}} + w_r\mathbf{e}_r, \quad (4)$$

where $\bar{\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}}{\alpha^+}$ and $w_r = \frac{\tau_r}{\alpha^+ + \tau_r}$. It is worth noting that each mean vector $\boldsymbol{\lambda}_r$ can be interpreted as a weighted average of a common barycenter, $\bar{\boldsymbol{\alpha}}$, and the r -th simplex vertex \mathbf{e}_r . In the light of the mixture structure of the EFD, its first order moment is straightforward:

$$\boldsymbol{\mu}_j = \mathbb{E}[Y_j] = \sum_{r=1}^D p_r \boldsymbol{\lambda}_{rj} = \bar{\alpha}_j \sum_r p_r (1 - w_r) + p_j w_j. \quad (5)$$

Though, the new parameterization based on $\boldsymbol{\mu}_j$, p_j and w_j ($j = 1, \dots, D$) is not variation independent, which may be more appropriate for inferential purposes, and in particular for Bayesian inference through MC methods. This is due to the existence of the following constraints:

$$0 < \bar{\alpha}_j = \frac{\boldsymbol{\mu}_j - p_j w_j}{1 - \sum_r p_r w_r} < 1 \quad j = 1, \dots, D. \quad (6)$$

In particular, from (6), we get the following inequalities referred to w_j , $j = 1, \dots, D$:

$$(i) \ w_j < \frac{\boldsymbol{\mu}_j}{p_j}, \quad (ii) \ w_j > \frac{\boldsymbol{\mu}_j}{p_j} - \frac{1 - \sum_r p_r w_r}{p_j}.$$

Since $0 < w_r < 1$ for every r , $0 < \sum_r p_r w_r < 1$, $p_j w_j < \boldsymbol{\mu}_j$ for every j , and $\sum_{r \neq j} p_r w_r < 1 - \boldsymbol{\mu}_j$, by summing up the two inequalities in (i) and (ii) we obtain the normalized version of w_j that is equal to:

$$\tilde{w}_j = \frac{w_j}{\min\left\{\frac{\boldsymbol{\mu}_j}{p_j}, 1\right\}}, \quad j = 1, \dots, D. \quad (7)$$

Finally, the parameterization of the EFD distribution depending on $\boldsymbol{\mu} \in \mathcal{S}^D$, $\mathbf{p} \in \mathcal{S}^D$, $\tilde{w}_j \in (0, 1)$ for every j , and $\alpha^+ > 0$ is variation independent since the parameters are not linked by any constraint.

In the light of this new parameterization, which explicitly includes the mean vector $\boldsymbol{\mu}$, we can now derive a regression model based on the EFD distribution. Let us denote the response matrix and the design matrix by $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)^\top$ and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ respectively. For each $i = 1, \dots, n$, the \mathbf{Y}_i are D -dimensional vectors on the simplex and \mathbf{x}_i are $(K + 1)$ -dimensional vectors.

Following a GLM strategy [8], the objective of the regression is the mean vector of \mathbf{Y}_i that is, in general terms, expressed by \mathbf{v}_i . To account for the constraints on the latter, namely positivity and sum to 1, we adopt a multinomial logit link function [7] as follows:

$$g(\mathbf{v}_{ij}) = \log \left(\frac{v_{ij}}{v_{iD}} \right) = \mathbf{x}_i^\top \boldsymbol{\beta}_j, \quad (8)$$

where $v_{ij} = \mathbf{E}(Y_{ij})$, $\mathbf{x}_i = (1, x_{i1}, \dots, x_{iK})^\top$ is the vector of covariates, and $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jK})^\top$ is a vector of regression coefficients. It is worth noting that D is fixed as the baseline category, therefore $\beta_{Dk} = 0$ for $k = 0, 1, \dots, K$. As a consequence, we get:

$$v_{ij} = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}_j) = \begin{cases} \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_j)}{1 + \sum_{r=1}^{D-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_r)}, & \text{for } j = 1, \dots, D-1 \\ \frac{1}{1 + \sum_{r=1}^{D-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_r)}, & \text{for } j = D. \end{cases} \quad (9)$$

In case \mathbf{Y}_i are Dirichlet distributed (therefore $v_{ij} = \bar{\alpha}_{ij}$) we recover the DirReg model [6], while if they are EFD distributed (so that $v_{ij} = \mu_{ij}$) we get the Extended Flexible Dirichlet regression (EFDRReg) model. Note that only the latter model enables the construction of D further group regression curves given by (4).

Inferential issues for the EFDRReg model are dealt with a Bayesian approach. Because of the mixture structure of the EFD, likelihood-based inferential approaches would be cumbersome from a computational and analytical point of view. Instead, from a Bayesian perspective, the finite mixture structure of the EFD distribution is advantageous in that it could be dealt with as an incomplete data problem [4]. We adopt the Hamiltonian Monte Carlo (HMC) [10] algorithm that is a generalization of the Metropolis algorithm which combines Markov Chain Monte Carlo (MCMC) and deterministic simulation methods. The HMC algorithm is implemented in the Stan modeling language through the standard No-U-Turn Sampler [12].

To generate the (simulated) posterior distributions for the unknown parameters, prior distributions must be specified. With regard to priors elicitation, we propose to take advantage of non- or weakly informative priors to induce the minimum impact on the posteriors [2]. We set a multivariate normal with zero mean vector and diagonal covariance matrix with “large” values of the variances as non-informative prior for the regression parameters $\boldsymbol{\beta}_j$. Furthermore, we select a Uniform(0, 1) prior for \tilde{w}_j , $j = 1, \dots, D$, and a Dirichlet prior with a hyperparameter of $\mathbf{1}$ for the vector \mathbf{p} . Last, we adopt a gamma prior Gamma(g, g) (with g assuming different “small” values) for the precision parameter α^+ .

3 An application to the Italian elections data

We compare the DirReg and the EFDRReg models through an application regarding the results of the Italian elections held on 4 March 2018 (data are freely available from the Italian Ministry of Interior’s webpage). The multivariate response is a 3-

dimensional vector composed by the percentages of votes got by the “Movimento 5 Stelle” (M5S), “Lega”, and “Other parties”. Data are collected on the 231 electoral districts into which the Italian territory is organized. We evaluate the impact of a continuous covariate concerning the level of “Employment Rate” (ER) that is defined as the ratio of the number of employed persons (aged 15-64) to the number of persons (aged 15-64). We define a regression model for the mean of the response vector as in (8) under the DirReg ($v_{ij} = \tilde{\alpha}_{ij}$) and the EFDReg ($v_{ij} = \mu_{ij}$) having fixed “Other parties” as the baseline category. For both models, we simulate a MCMC of length 10 000 —having set a burn-in period of 5 000 —and we check the convergence to the stationary posterior distribution both via graphical evaluation and diagnostic indices.

Table 1 shows the posterior means, standard deviations, and 95% credible intervals (CIs) of the parameters of the competing models. The regression coefficients are similar for both the models and the ER covariate has a mild-negative significant association with the percentage of votes got by the M5S whereas it has a mild-positive significant association with the percentage of votes got by the Lega party. The goodness of fit of the two models is compared through the Watanabe-Akaike information criterion (WAIC) [5]. The EFDReg provides a much better fit than the standard DirReg, since its WAIC value (-1486.5) is smaller than the DirReg value (-1368.2). Figure 1 shows the fitted curves under the DirReg and EFDReg models and the component means λ_r ($r = 1, 2, 3$) of the latter. Note that the third mixture component (orange curves) represents a small group (posterior mean of p_3 equal to 0.05) corresponding to outlier observations. Clearly a deeper study of the model and, in particular, of the form of the group regression models is necessary, though these first results show the feasibility and the good fit abilities of the new model.

Model	Parameter	Post. Mean	Post. SD	2.5%	97.5%
DirReg	$\beta_{0,1}$	1.64	0.13	1.40	1.88
	$\beta_{1,1}$	-0.04	< 0.001	-0.04	-0.03
	$\beta_{0,2}$	-4.09	0.20	-4.48	-3.72
	$\beta_{1,2}$	0.05	< 0.001	0.04	0.06
	α^+	47.31	3.00	41.62	53.22
EFDReg	$\beta_{0,1}$	1.45	0.09	1.26	1.63
	$\beta_{1,1}$	-0.03	< 0.001	-0.04	-0.03
	$\beta_{0,2}$	-4.74	0.24	-5.20	-4.26
	$\beta_{1,2}$	0.06	< 0.001	0.05	0.07
	α^+	91.65	5.30	80.02	99.53
	p_1	0.41	0.09	0.25	0.58
	p_2	0.55	0.09	0.37	0.71
	p_3	0.05	0.01	0.02	0.08
	\tilde{w}_1	0.03	0.02	0.00	0.07
	\tilde{w}_2	0.31	0.05	0.22	0.40
	\tilde{w}_3	0.40	0.04	0.33	0.47

Table 1 Posterior means, standard deviations (SDs), and 95% CIs for the parameters under the DirReg and the EFDReg models.

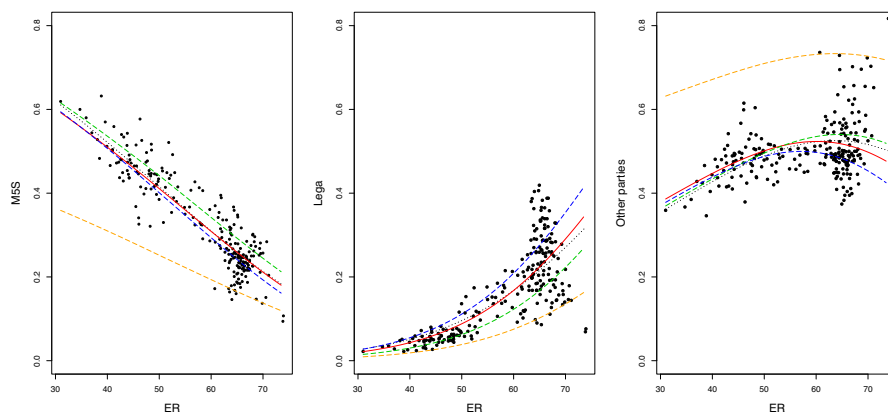


Fig. 1 Fitted regression curves under the DirReg (black dotted) and the EFDReg (red solid) models. Long dashed curves represent the component means λ_1 (green), λ_2 (blue), and λ_3 (orange) of the EFD.

References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. The Blackburn Press, London (2003)
2. Albert, J.: Bayesian computation with R. Springer Science & Business Media (2009)
3. Ascari, R.: A Family of Flexible Mixture Distributions for Constrained Data. Doctoral dissertation (2019)
4. Frühwirth-Schnatter, S.: Finite mixture and Markov switching models. Springer Science & Business Media (2006)
5. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis, 3rd edn. CRC Press, London (2013)
6. Hijazi, R.H., Jernigan, R.W.: Modelling compositional data using Dirichlet regression models. *J. Appl. Probab. Statist.* **4**, 77-91 (2009)
7. Maier, M.J.: Dirichletreg: Dirichlet regression for compositional data in R. Research Report Series, Department of Statistics and Mathematics, University of Economics and Business, Vienna (2014)
8. McCullagh, P., Nelder, J.: Generalized linear models. Chapman & Hall, London (1989)
9. Migliorati, S., Ongaro, A.: The Extended Flexible Dirichlet Model: Some Theoretical and Computational Issues. In *Applied Stochastic Models and Data Analysis (ASMDA)*. Piraeus, Greece (2015)
10. Neal, R.M.: An Improved Acceptance Procedure for the Hybrid Monte Carlo Algorithm. *J. Comput. Phys.*, **111**(1), 194-203 (1994)
11. Ongaro, A., Migliorati, S.: A Dirichlet Mixture Model for Compositions Allowing for Dependence on the Size. In *Advances in Latent Variables*, 101-11 (2014)
12. Stan Development Team: Stan Modeling Language Users Guide and Reference Manual (2016) <http://mc-stan.org/>

Turning big data into smart data: two examples based on the analysis of the Mappa dei Rischi dei Comuni Italiani

Trasformare i big data in smart data: due esempi di analisi della Mappa dei Rischi dei Comuni Italiani

Oleksandr Didkovskyi, Alessandra Menafoglio, Piercesare Secchi, Giovanni Azzone

Abstract The recently presented *Mappa dei Rischi dei Comuni Italiani* is a freely accessible web portal, implemented by ISTAT, which provides integrated data on different natural risks in Italian municipalities together with socio-economic and demographic data. We here illustrate two paradigmatic examples where the big data of the *Mappa* are transformed into smart data using advanced methods for descriptive statistics, thus providing interesting insights on local patterns and regional trends in terms of building stock vulnerability and social and material vulnerability.

Abstract *La Mappa dei Rischi dei Comuni Italiani è un portale web pubblico, di recente reso operativo da ISTAT; per ogni comune italiano esso fornisce in modo integrato dati relativi ai rischi naturali insieme ad indicatori socio-economici e demografici. In questo lavoro illustriamo due esempi di analisi nelle quali i big data della Mappa sono trasformati in smart data utilizzando approcci avanzati di statistica descrittiva. Gli esempi forniscono interessanti visioni, sia a livello locale che nazionale, sulla vulnerabilità del patrimonio edilizio italiano e sulla vulnerabilità sociale e materiale del paese.*

Key words: Object oriented data analysis, compositional data, high-dimensional descriptive statistics

Oleksandr Didkovskyi
MOX, Department of Mathematics, Politecnico di Milano *and*
Center for Analysis Decision and Society, Human Technopole, Milano.
e-mail: oleksandr.didkovskyi@polimi.it

Alessandra Menafoglio
MOX, Department of Mathematics, Politecnico di Milano.
e-mail: alessandra.menafoglio@polimi.it

Piercesare Secchi
MOX, Department of Mathematics, Politecnico di Milano *and*
Center for Analysis Decision and Society, Human Technopole, Milano.
e-mail: piercesare.secchi@polimi.it

Giovanni Azzone
Department of Management, Economics and Industrial Engineering, Politecnico di Milano.
e-mail: giovanni.azzone@polimi.it

1 Introduction

On the 18th February 2019 the Department Casa Italia of the Italian Government presented to the public the *Mappa dei Rischi dei Comuni Italiani*¹ (MRCI). The freely accessible web portal, implemented by ISTAT, provides integrated information on different natural risks in Italian municipalities - such as earthquake, flooding, landslide, volcano eruption - in conjunction with socio-economic and demographic data. It offers the possibility of viewing and downloading indicators, charts and maps, together with guided interactive features for data searching and filtering.

The Casa Italia task force² was established in 2016 to develop a plan for housing and land care aimed at better protection of citizens, and public and private goods. The goal was to define the constituent elements of a national policy for the promotion of housing safety. Quality of living was identified as of primary importance for the mission of Casa Italia, with a particular emphasis on policies for the promotion of security of residential buildings against natural risks. The key idea was that of a multi-hazard approach to risk, focusing on the security of places where people live which are instrumental to their individual safety. As part of the proposed action plans, the MRCI aims to create a widespread awareness of the fragility of the Italian territory. Despite fragmentation, dispersion and diversity of the available databases, an information platform was built to allow a homogeneous and integrated view of the natural risks within the Italian territory.

In the first stage of the project, Casa Italia aimed at integrating and enhancing the rich information on natural risks already available as the result of numerous and intense research activities carried on by several local and national institutions (ISTAT, INGV, ISPRA, ENEA, CNR, MIBACT). Part of this action was devoted to identify the sources of information and the data bases allowing for a unified and integrated vision of the natural risks insisting on the Italian territory, with particular reference to the three factors that compose risk, i.e., hazard, vulnerability, exposure. Due to the mission of Casa Italia, the survey was limited to databases that (a) were elaborated by official and national research institutes, (b) had coverage of the entire national territory, and (c) had a spatial resolution sufficient to allow identification and comparison of local specificities. With reference to the latter point (c), the municipality (i.e., *Comune*) was identified as the smallest spatial statistical unit for the actual analyses. This choice is motivated by the need of integrating and fusing data from different sources, with different spatial resolutions, and originally generated for different aims. For instance, at the national level coverage, the municipality represents

¹ <http://www4.istat.it/it/mappa-rischi>

² The task force 'Casa Italia' of the Italian Presidency of the Council of Ministers was established on September 23, 2016. Its members were Giovanni Azzone (Project Manager and Scientific Director), Massimo Alvisi, Michela Arnaboldi, Alessandro Balducci, Marco Cammelli, Guido Corso, Francesco Curci, Daniela De Leo, Carlo Doglioni, Andrea Flori, Manuela Grecchi, Massimo Livi Bacci, Maurizio Milan, Alessandra Menafoglio, Pietro Petrarola, Fabio Pammolli, Davide Rampello, Piercesare Secchi. The 3rd of July 2017, the Italian Presidency of the Council of Ministers established 'Casa Italia' as one of its departments, committed to the prevention against natural risks (<http://www.casaitalia.governo.it/it/>). The authors acknowledge the task force 'Casa Italia' for the scientific discussions that were inspirational to the present work.

the smallest administrative unit for which aggregated data on the vulnerability of residential buildings are today available, and these data come from the last national census of 2011. The dataset consists of observations taken at 7983 Municipalities, updated at 2018. Additionally, information about the same indices aggregated by provinces and regions is also provided.

To illustrate the richness of MRCI, in this paper we report on two analyses, which we consider as paradigmatic examples of the application of non trivial statistical descriptive methods and algorithms aimed at transforming big data into smart data, which are then made openly available to policy makers and the citizens.

The first analysis was developed as part of the Casa Italia project [3] and regards the age of the Italian building stock. For each Italian municipality, MRCI reports the distribution of the age of the buildings, grouped in $p = 9$ non overlapping time intervals. These object data are compositional [1]; for their analysis we should embed them in a proper space, which we take to be the simplex in R^8 endowed with the Aitchison geometry [1]. For the sake of this illustration and ease of visualization, we shall represent these data when the original classes are aggregated into 3 classes of ages, and thus represent the data in the simplex embedded in R^2 . Within this space we explore data variability by performing a suitable PCA [1]; the projection of the age distributions along the main directions of their variability as object data, offers a powerful representation for the understanding of age of the Italian building stock over the entire national territory.

The second example considers the *Indice di Vulnerabilità Sociale e Materiale*³ (IVSM) which is a quality of life index computed by ISTAT for each Italian municipality and reported in MRCI. For this analysis we consider the Italian provinces as statistical units. For each province we look at the distribution of IVSM among the municipalities it is composed of. Location and scale of these IVSM distributions – not to mention their quantiles – are of course of great interest to the policy maker, indicating respectively the degree of social and material vulnerability of the province and its heterogeneity among the municipalities. Hence an exploration of the IVSM distributions which treats each of them as a whole data object seems appropriate, as opposed to a more naive approach which would work separately on their mean, their standard deviations or their quantiles. The final goal of the analysis is to cluster provinces according to similarity of the distribution of IVSM among their municipalities. This will offer to the policy maker, at the national level, a picture of the country in terms of homogeneous macroregions, characterized by similar issues in terms of social and material vulnerability. In the next two sections we illustrate the two examples. Final comments will close the paper.

2 Compositional analysis of the age of the building stock

The age of a building is a key element to determine its vulnerability to seismic events, as it is directly associated with the seismic regulations in force at the time of the construction, as well as with the advance of building technologies. For the purpose of this illustration, we shall consider the distribution – within municipalities

³ <http://ottomilacensus.istat.it/documentazione/>

– of the age of the building stock in the following three classes: before 1919, between 1919 and 1980, after 1980. These classes are representative, respectively, of extremely old buildings, buildings of medium age, and buildings erected under the more recent regulations in terms of seismic risk (for further details, see [3]). These data are compositional in their nature, they belong to the simplex embedded in R^2 which we endow with the Aitchison geometry [1]. Accordingly, we perform a compositional principal component analysis (PCA), and we then explore the directions of the simplex along which the dataset displays its maximum variability.

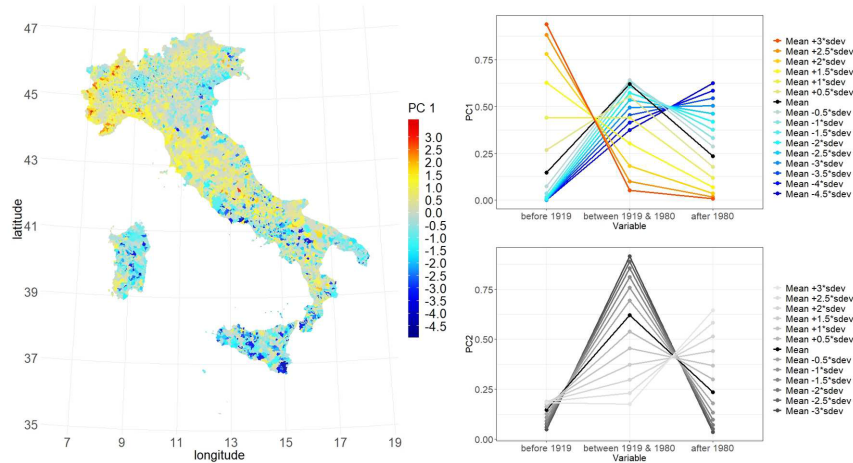


Fig. 1 Standardized scores along first compositional PC of the distribution of building stock in terms of age (left) and variation of the compositions along the first and second compositional PCs.

The left panel of Fig. 1 displays the scores along the main mode of variability of the dataset, which captures 86% of the total variability. The right panel of Fig. 1 displays the variation of the composition of the building stock when moving along the first PC (top panel) and the second PC (bottom panel), to support interpretations. High-scores along the first PC are representative of municipalities with a prevalence of very old buildings over the most recent ones and viceversa. Spatial patterns of high scores are particularly visible along the Apennines in central Italy, in Liguria and in the outer parts of Piemonte, as well as in some parts of Friuli Venezia Giulia. The score values along the first PC thus provide an indication of the overall degree of vulnerability of the building stock with respect to the Italian mean, and suggest to the policy maker the areas that may deserve specific interventions for the promotion of security of buildings against seismic risk.

3 Symbolic analysis of IVSM

MRCI reports different indicators on the vulnerability and resilience of Italian municipalities, taken from the socio-economic perspective. We here focus on IVSM,

which is given at the scale of municipality, and ranges in $[90,120]$, higher values being associated with higher vulnerability. Although data in MRCI are at the scale of municipality, we consider their aggregation at a lower level of resolution – the province – to provide views of the local social vulnerability, as well as of the homogeneity of the indicators on the administrative unit; clearly, other aggregations would be possible (e.g., regional). These multi-scale views are relevant and are made possible by the MRCI. They allow the citizen, as well as the decision maker, to perform analysis and evaluations at a micro-scale or at an aggregated scale. Having grouped the micro-data according to the province, we consider as data object the probability density function (PDF) of IVSM within the province, as estimated by kernel smoothing. The location of the PDF provides indications of the overall vulnerability of the area, whereas the scale of the PDF indicates the regional homogeneity. We here aim to identify cluster of provinces characterized by a similar degree of vulnerability and homogeneity. The analysis and clustering of PDF data requires to properly define a notion of similarity between the PDFs of IVSM estimated at different provinces. As in the case discussed in Section 2, a Euclidean metric (or L^2 metric) would not correctly represent the data constraints; instead, metrics for distributional data are more appropriate and should be preferred. The Wasserstein metric is widely used in Symbolic Data Analysis (SDA, [2]) as a measure of dissimilarity among distributional data, and has insightful interpretation in terms of optimal transport. As an alternative, a functional Aitchison geometry (a.k.a., Bayes Hilbert geometry, [4]) may be used instead. For the sake of brevity, in this illustration we shall focus on the former metric only.

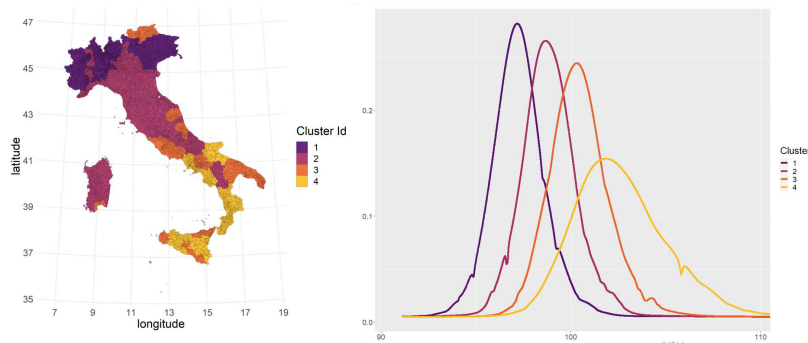


Fig. 2 Map of Italy where provinces are colored according the associated cluster (left) and centroids of the clusters of the IVSM PDFs (right).

Fig.2 reports the results of hierarchical clustering based on a Wasserstein metric with Ward linkage, having set to 4 the number of clusters. The left panel of Fig.2 reports a map of the identified clusters, whereas the right panel reports the centroids of the cluster, computed as Frechét mean of the PDFs within the clusters. These results show that the four clusters differ for both location and scale. The first cluster

– associated with some of the Northern provinces – is characterized by an overall low social and material vulnerability and higher regional homogeneity. Moving to the following clusters, one may observe increasing vulnerability and regional heterogeneity; a North-South trend is clearly visible. The province of Bolzano stands in clear contrast with the neighboring provinces, suggesting a local outlyingness. These analysis can support national and local administrators to evaluate the social and material vulnerability of the population and its possible resilience in response to a natural event.

4 Conclusion and further discussion

MRCI provides a rich and integrated framework to allow for multi-scale analyses on the natural and social risks insisting on the Italian territory. We here embrace the viewpoint of object oriented data analysis ([5]), and recognize as key elements of the analysis the identification of the *data object*, the choice of its geographical scale of reference and of the most appropriate geometry for its mathematical representation – choices that should indeed be guided by the goal of the analysis. In fact, turning *big data* into *smart data* inherently requires a strong interplay of advanced statistical method and experts' knowledge, to provide meaningful summaries and insights, and particularly to identify the *object* and the *objective* of the analysis, which still remain of primary importance even in the data deluge era. As paradigmatic examples of this methodological approach, we illustrated two object oriented analyses, where the use of advanced statistical method for aggregated data served the purpose of getting insights on local patterns and regional trends. These analyses may be extended to provide further views and predictive models, with the final aim of increasing the understanding and the awareness of the complex and multi-scale character of the social and natural risks insisting over the country, and, ultimately, to support the decision making of citizens, local administrators and policy makers.

References

1. Pawlowsky-Glahn, V., Buccianti, A., Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Chichester (2011).
2. Billard, L., Diday, E., Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons, Chichester (2007).
3. Presidenza del Consiglio dei Ministri, Struttura di Missione Casa Italia, Rapporto sulla Promozione della sicurezza dai Rischi naturali del Patrimonio abitativo, <http://www.casaitalia.governo.it/it/approfondimenti/rapporto-sulla-promozione-della-sicurezza/> (2017).
4. Van den Boogaart, K. G., Egozcue, J., Pawlowsky-Glahn, V. Bayes Hilbert spaces. Aust. N. Z. J. Stat., Vol. 56, No. 2, 171–194 (2014).
5. Marron, J. S., Alonso, A. M.: Overview of object oriented data analysis. Biometrical Journal, Vol. 56, No. 5, 732–753 (2014).

Hidden Markov Model estimation via Particle Gibbs

Stima di Hidden Markov Model tramite Particle Gibbs

Pierfrancesco Alaimo Di Loro, Enrico Ciminello and Luca Tardella

Abstract When the likelihood of the model is not explicitly available, standard Markov Chain Monte Carlo techniques may become impractical. This paper has the aim to investigate a combination of Sequential Monte Carlo and Metropolis-Hastings algorithm in the spirit of the pseudo-marginal approach. This produces algorithms known as Particle MCMC which are part of a powerful and flexible class of algorithms called Exact-Approximate MCMC. They establish a new paradigm in parameter estimation in the non-linear and non-Gaussian Hidden Markov Models (HMM). In this paper, the Particle Gibbs sampler is used to recover the parameter of an HMM applied to the time series of worldwide annual earthquakes of magnitude 7 or greater occurred in the 21st century.

Abstract *Quando la verosimiglianza non è trattabile analiticamente, allora le tecniche Markov Chain Monte Carlo standard possono risultare inattuabili. Questo articolo ha l'obiettivo di analizzare una combinazione di tecniche Sequential Monte Carlo e dell'algoritmo Metropolis Hastings alla luce dell'approccio pseudo-marginale. Tale combinazione produce algoritmi Particle MCMC, parte di una classe molto flessibile nota come Exact-Approximate MCMC, che costituiscono un nuovo paradigma per la stima nell'ambito di Hidden Markov Model non lineari e non Gaussiani. In questo articolo è proposto l'utilizzo del Particle Gibbs per la stima di un modello HMM applicato alla serie annuale dei terremoti di magnitudo 7 o superiore avvenuti su scala mondiale durante il XXI secolo.*

Key words: MCMC, HMM, Bayesian, particle filters, time series, earthquake

Pierfrancesco Alaimo Di Loro

“La Sapienza” University of Rome, Piazzale Aldo Moro, 5, Rome, e-mail: pierfrancesco.alaimodiloro@uniroma1.it

Enrico Ciminello

“La Sapienza” University of Rome, Piazzale Aldo Moro, 5, Rome, e-mail: enrico.ciminello@uniroma1.it

Luca Tardella

“La Sapienza” University of Rome, Piazzale Aldo Moro, 5, Rome, e-mail: luca.tardella@uniroma1.it

Introduction

Interest in a process which can only be observed indirectly is a problem encountered in a variety of applications: biological sequences analysis [4], speech recognition [15] and time series analysis in general. Sometimes, the presence of a latent process depends on the theoretical framework, but often it is introduced for convenience.

Consider, for example, a Bayesian framework where the parameter of interest $\theta \in \Theta$ has a posterior density $p_y(\theta)$ that is not analytically available. The introduction of a latent variable $z \in \mathcal{Z}$ usually allows for an easier formulation and manipulation of the model. A typical estimation technique in such contexts is based on the *Gibbs Sampler* which samples alternatively from the conditionals $p_y(\theta|z)$ e $p_y(z|\theta)$. This sampling scheme can very often ease programming and lead to elegant algorithms. On the other hand, if $p_y(\theta)$ were known analytically or cheap to compute, it would often be possible to generate “more efficient” samples $\{\theta_i\}$ from a Markov chain by means of a classic *Metropolis-Hastings* (MH) algorithm. This led to the development of MCMC algorithms that try to combine possible computational efficiency of sampling directly from $p_y(\theta)$ and implementational ease of augmented schemes.

Often the likelihood, even if not analytically available, can be estimated in an unbiased way using Monte Carlo methods: this is the case of *Hidden Markov Models* (HMM, introduced in Section 1). *Sequential Monte Carlo* (SMC) allows to simulate from the unobservable process and get unbiased estimate of the needed density, paving the way to the application of approximated algorithms.

Section 2 focuses on *Particle MCMC* (PMCMC), introduced in [2], which enables to perform parameter estimation for HMMs. These algorithms combine the ability of SMC to provide an unbiased estimate of the marginal likelihood of the process along with the pseudo-marginal approach [1], leading to a procedure that targets the true joint distribution of the parameters and the latent process.

In Section 3, a non-linear and non-Gaussian HMM is considered to model the annual time series of earthquakes of magnitude 7 or greater occurred worldwide during the 21st century. The Particle Gibbs sampler (PG), *Gibbs-style* version of the PMCMC, is used in order to recover the posterior distribution of the parameters and the resulting estimates are discussed in comparison to the ones obtained by [3] on the same data using the Particle Marginal Metropolis Hastings.

1 Hidden Markov Models

[6] describe *Hidden Markov Models* (HMM) as *the most successful statistical modeling ideas that have come up in the last forty years*. The partial bibliography by [5] gives a partial account of the wide scope of the domain: speech recognition [10], econometrics [12], computational biology [13], etc. The use of the hidden states makes such models generic enough to handle a variety of complex real-world time

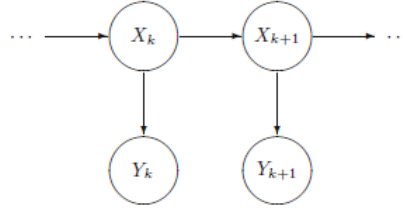


Fig. 1 *Hidden Markov Model* graph representation. Circles represent variables and arrows represent dependency relationships.

series, while the relatively simple formulation still allows for the use of efficient computational procedures.

An HMM is a process composed of a Markov Chain $\{X_t\}_{t=1}^T \subset \mathcal{X}^T$ with initial density $X_1 \sim p_\theta(\cdot)$ and transition probability $X_{t+1}|(X_t = x) \sim f_\theta(\cdot|x)$, where $\theta \in \Theta$ lives on a space of arbitrary dimension. This process $\{X_t\}$ is not observed directly, but through another stochastic process $\{Y_t\}_{t=1}^T \subset \mathcal{Y}^T$, whose observations are assumed to be independent conditionally on X_1, X_2, \dots, X_T and to have conditional density $Y_t|(X_t = x_t)_{t=1}^T \sim p_\theta(\cdot|x_t)$. The dependence structure of an HMM can be represented by a graphical model as in Figure 1, where nodes (circles) in the graph correspond to the random variables, and the edges (arrows) represent the structure of the joint probability distribution.

When the space \mathcal{X} of the hidden state X_t is discrete, the likelihood can be computed analytically while, if the space of X_t is continuous, it can be computed only when the models for X_t and Y_t are linear and Gaussian [11]. This stimulated the interest in alternative strategies that could be applied to more general frameworks. [9] proposed a first attempt of approximating the target distribution using a sequential version of the Monte Carlo importance sampling known as *Sequential Monte Carlo* (SMC). It is based on a recursive filtering approach, so that the received data can be processed sequentially rather than as a batch. Anyway, the choice of a suitable importance distribution in the form $q(x_{1:t}|y_{1:t})$ which is easy to sample from is not trivial. The procedure may be eased by the use of an auxiliary distribution that can be factored as follows:

$$q(x_{1:t}|y_{1:t}) = q(x_1|y_1) \prod_{i=1}^t q(x_i|x_{1:i-1}, y_{1:i}).$$

The sampling problem reduces to the one of recursively sample an arbitrarily large number N of *particles* (latent process samples) from univariate distributions of the form above. Approximations to the marginal likelihood and posterior densities of either parameters and latent states can be obtained from the set of the N weighted particles, where the weights are proportional to the conditional likelihood $p_\theta(y_t|x_t)$.

2 Particle Markov Chain Monte-Carlo

PMCMC methods rely on a non-trivial and non-standard combination of MCMC and Sequential Monte Carlo methods, which takes advantage of the strength of both components. MCMC machineries requires either $p_\theta(\mathbf{y})$ to be analytically tractable or $p_\theta(\mathbf{z}|\mathbf{y})$ to be sampled from. In the case of non-linear and non-gaussian HMMs, SMC may be used to produce likelihood estimates or to sample from $p_\theta(\mathbf{z}|\mathbf{y})$ in order to yield flexible algorithms, which approximate the exact ones in the spirit of the pseudo-marginal approach [1]. PMCMC can be used to solve the same inferential problems of SMC and its extensions (IBIS by [7]; SMC² by [8]) but, in spite of its reliance on SMC methods, usually it is *much more robust and less likely to suffer the depletion problem* [2].

The Particle marginal Metropolis-Hastings

Recalling the standard decomposition for the posterior density $p(\theta, x_{1:T}|\mathbf{y}_{1:T}) = p(\theta|\mathbf{y}_{1:T})p_\theta(x_{1:T}|\mathbf{y}_{1:T})$ it would be natural to suggest for the MH update a proposal of the form $q(\theta^*, x_{1:T}^*|\theta, x_{1:T}) = q(\theta^*|\theta)p_{\theta^*}(x_{1:T}^*|\mathbf{y}_{1:T})$, for which the proposed $x_{1:T}^*$ is perfectly adapted to the proposed θ^* , that is sampled from an arbitrarily chosen proposal distribution $q(\cdot|\theta)$. The resulting MH acceptance ratio depends on the marginal likelihood $p_\theta(\mathbf{y}_{1:T})$ that, in an HMM context, may be replaced by its SMC counterpart $\hat{p}_\theta(\mathbf{y}_{1:T})$. This leads to the *Particle Marginal Metropolis Hastings* (PMMH).

Particle Marginal MH

1. **Initialization.** Choose $\theta^{(0)}$ and run the SMC. Compute $\hat{p}_{\theta^{(0)}}(\mathbf{y}_{1:T})$ and sample $x_{1:T}^{(0)}$ from $\{x_{1:T}^{(i)}\}_{i=1}^N$ with weights $\{\omega_T^{(i)}\}_{i=1}^N$.
2. **Recursions.** For each $j \in \{1, \dots, M\}$:
 - a. propose θ^* from $q(\cdot|\theta^{(j-1)})$;
 - b. run the SMC to get $\hat{p}_{\theta^*}(\mathbf{y}_{1:T})$ and sample a new hidden path $x_{1:T}^*$;
 - c. accept θ^* and $x_{1:T}^*$ in the chain $(\theta^{(j)}, x_{1:T}^{(j)})$ with probability:

$$\alpha = 1 \wedge \frac{p(\theta^*)\hat{p}_{\theta^*}(\mathbf{y}_{1:T})q(\theta^{(j-1)}|\theta^*)}{p(\theta^{(j-1)})\hat{p}_{\theta^{(j-1)}}(\mathbf{y}_{1:T})q(\theta^*|\theta^{(j-1)})}.$$

Set $\theta^{(j)} = \theta^{(j-1)}$ and $x_{1:T}^{(j)} = x_{1:T}^{(j-1)}$ otherwise.

The PMMH update leaves $p(\theta, x_{1:T}|\mathbf{y}_{1:T})$ invariant and, under weak assumptions, ergodic [2]. The goodness of the resulting chain depends only on the more or less accurate choice of the proposal density. The problem is that such a choice is theoretically arbitrary, and there is a really low number of guidance on its appropriate formulation.

The Particle Gibbs

The *Particle Gibbs* consists of using a *Gibbs-style* update to sample iteratively $\theta \sim p(\theta|x_{1:T}, y_{1:T})$ and $X_{1:T} \sim p_\theta(x_{1:T}|y_{1:T})$. It is often possible to easily sample from $p(\theta|x_{1:T}, y_{1:T})$, and thus the issue of designing a proposal density for θ is encompassed. Sampling directly from $p_\theta(x_{1:T}|y_{1:T})$ is usually not feasible and, when replacing it by sampling from the SMC approximation $\hat{p}_\theta(x_{1:T}|y_{1:T})$, the convergence properties of the *Gibbs Sampler* (GS) to the target density $p_{x_{1:T}}(\theta|y_{1:T})$ do not hold anymore. A valid particle approximation to the GS requires the use of a special type of PMCMC update, called the *conditional SMC* update [2].

Particle Gibbs

1. **Initialization.** Set the arbitrary starting points $\theta^{(0)}$, $x_{1:T}^{(0)}$ and $B_{1:T}^{(0)}$.
2. **Recursions.** For $j \in \{1, \dots, M\}$:
 - a. sample $\theta^{(j)} \sim p(\cdot|x_{1:T}^{(j-1)}, y_{1:T})$;
 - b. sample $x_{1:T}^{(j)}$ running a conditional SMC.

Stationarity and ergodicity of the resulting chain with respect to the target density are ensured in [2].

3 Applications

The earthquakes data used in [3] has been already analyzed by [14]. It describes the number of annual earthquakes with a magnitude of 7 or over (on the Richter scale) occurred worldwide along the 21st century. Usually this kind of data would be modeled using the Poisson distribution. However, the series is affected by over-dispersion and, furthermore, it presents significantly positive auto-correlations making unrealistic any hypothesis of independence. These features of the data motivated [14] to use a time-dependent parameter in the Poisson that may be resumed as follows:

$$\begin{cases} Y_t | (X_t = x_t) \sim \text{Pois}(\exp(\gamma + x_t)) \\ X_t | (X_{t-1} = x_{t-1}) \sim N(\phi x_{t-1}, \tau^2). \end{cases}$$

[3] applied the PMMH to get Bayesian estimates of the parameters: the results are exposed in Table 3. However, the principal drawback of the PMMH application is that the choice of the proposal for the random walk may play a key role in determining the quality of the final chains. [3] themselves were aware that better results could be achieved by a more careful choice of the proposal density, allowing for a better exploration of the parameters' space.

An alternative solution would be the application of a PG sampler in place of the PMMH as it does not require the design of any proposal distribution. The full conditionals of the parameters can be easily evaluated and they all turn out to belong to known parametric families. The obtained estimates are presented along with the

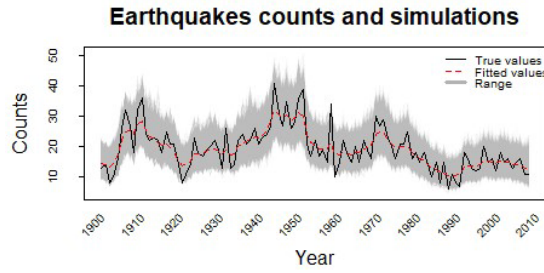


Fig. 2 Comparison of the observed counts with the PMCMC marginal predictive simulated trajectories and the average estimate.

Table 1 Estimates and ESS using PMMH [3] and PG.

Particle Marginal MH			Particle Gibbs		
Parameters	Mean	ESS	Parameters	Mean	ESS
ϕ	0.88	640	ϕ	0.897	1232
τ^2	0.02	683	τ^2	0.023	936
γ	2.93	442	γ	2.783	555

effective sample sizes of the chains in Table 3. They are really close to the ones by [3] but outplay them in terms of *Effective Sample Size (ESS)*. Finally the average trajectory, plotted against the observed counts, is presented in Figure 2 along with the whole range of estimated paths. This plot shows relatively good fit with the series of counts totally contained in the posterior predictive simulations range.

4 Concluding remarks

The aim of this work was to introduce the Particle MCMC methodology and to provide a comparison between the PMMH and the PG algorithm. In the analyzed application, if the random walk of the PMMH were implemented with an independent proposal, then the chains of the parameters would exhibit stickiness, leading to non satisfactory behaviours. [3] had to perform an accurate tuning of the covariance matrix of the proposal density in order to achieve well-behaved chains, without any guarantee that results could have been further improved.

The application of the PG spares the tuning and provides good results, characterized by fast convergence and good mixing of the chains. We may conclude that, whenever the model is simple enough and the correlation between parameters and latent states is not strong, the application of the PG is encouraged as its implementation is straightforward and estimation usually is more accurate.

References

- [1] Christophe Andrieu and Gareth O Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, pages 697–725, 2009.
- [2] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.
- [3] Jack Baker and Paul Fearnhead. Exact approximate Markov chain Monte Carlo. 2015.
- [4] Ewan Birney. Hidden Markov models in biological sequence analysis. *IBM Journal of Research and Development*, 45(3.4):449–454, 2001.
- [5] Olivier Cappé. Ten years of HMMs. URL: www.tsi.enst.fr/~cappe/docs/hmmbib.html, 2001.
- [6] Olivier Cappé, Eric Moulines, and Tobias Rydén. Inference in hidden markov models. In *Proceedings of EUSFLAT Conference*, pages 14–16, 2009.
- [7] Nicolas Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.
- [8] Nicolas Chopin, Pierre E Jacob, and Omiros Papaspiliopoulos. SMC2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):397–426, 2013.
- [9] JE Handschin. Monte Carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica*, 6(4):555–563, 1970.
- [10] Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.
- [11] Rudolph Emil Kalman et al. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.
- [12] Chang-Jin Kim, Charles R Nelson, et al. State-space models with regime switching: classical and Gibbs-sampling approaches with applications. *MIT Press Books*, 1, 1999.
- [13] Timo Koski. *Hidden Markov models for bioinformatics*, volume 2. Springer Science & Business Media, 2001.
- [14] Roland Langrock. Some applications of nonlinear and non-Gaussian state-space modelling by means of hidden Markov models. *Journal of Applied Statistics*, 38(12):2955–2970, 2011.
- [15] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

A note on marginal effects in logistic regression with independent covariates

Una nota sugli effetti marginali nella regressione logistica con covariate indipendenti

Marco Doretti

Abstract Starting from a logistic model with two independent and non-interacting covariates, in this paper the behaviour of marginal effects is studied. In detail, a setting with a continuous and a binary covariate is considered, the focus being on the effect of the continuous variable once the binary one is marginalized over. Building on some recent proposals in the literature, optimal bounds for the value of such an effect - which is not constant but depends on the covariate level - are derived. Also, it is shown that this marginal effect owns two interesting properties related to symmetry.

Abstract *A partire da un modello logistico con due covariate indipendenti senza interazione, questo articolo studia il comportamento degli effetti marginali. Nel dettaglio, si considera il caso in cui una covariata è continua e l'altra è dicotoma, ponendo l'attenzione sull'effetto della prima in seguito alla marginalizzazione rispetto alla seconda. Sviluppando alcuni risultati recentemente proposti in letteratura, vengono derivati i limiti ottimali per questo effetto, il quale non è costante ma dipende dal valore della covariata. In aggiunta, si mostra che questo effetto marginale possiede due interessanti proprietà legate alla simmetria.*

Key words: conditional effect, logistic regression, marginal effect

1 Introduction

Recently, the relationship between marginal and conditional effects in logistic regression has been studied. Specifically, given a logistic model for a binary outcome (Y) conditional on two variables (X and W), the exact formula for the marginal effect of X on Y , on the odds ratio scale, has been derived for the case where W is also binary; see [6]. This result can be thought of as an analogue for logistic regression

Marco Doretti
University of Perugia, e-mail: marco.doretti@unipg.it

of the well-known Cochran's formula [1] holding in the linear framework. In [6], different formulas are given depending on the nature of X , which can be either continuous or binary. In the former case, effects are expressed as derivatives, whereas in the latter they are expressed as difference in levels. When X is continuous, it is important to bear in mind that the functional form of the logistic model for Y given $X = x$ only is not linear in x ; see [3]. Therefore, the target marginal effect is not constant and generally depends on the value x .

In this work, the focus is on the setting where X and W are independent, in which the marginal and conditional effects of X on Y are different as a consequence of the well-known fact that odds-ratios are noncollapsible measures; see for example [5]. More specifically, it is assumed that X is continuous and not interacting with W in the model for Y . Within such a setting, the marginal effect derived in [6] takes a simple form and can be bounded in a very intuitive way, linking to (more general) results in the literature dealing with effect reversal [2] and bias due to omitted covariates in generalized linear models [4]. However, the naive bounds obtained by the formulas in [6] are not optimal, in the sense that other narrower bounds exist. In this paper, such optimal bounds for the marginal effect are derived and expressed as a function of the conditional logistic regression model parameters. Within this discussion, other appealing features concerning symmetry of the marginal effect, when regarded as a function of x , are also highlighted.

2 Optimal bounds for the marginal effect

Given two independent random variables X and W , with X continuous and W binary, acting as covariates in a logistic regression model without interaction

$$\log \frac{P(Y = 1 \mid X = x, W = w)}{P(Y = 0 \mid X = x, W = w)} = \beta_0 + \beta_x x + \beta_w w$$

for an outcome Y , results in [6] show that the marginal effect of X on Y on the logistic scale

$$\beta(x) := \frac{d}{dx} \log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)}$$

can be written as

$$\beta(x) = \beta_x \{1 - \Delta_y(x) \Delta_w(x)\}, \quad (1)$$

where the two probability differences

$$\begin{aligned} \Delta_y(x) &= P(Y = 1 \mid W = 1, X = x) - P(Y = 1 \mid W = 0, X = x) \\ \Delta_w(x) &= P(W = 1 \mid Y = 1, X = x) - P(W = 1 \mid Y = 0, X = x) \end{aligned}$$

have the same sign so that the term in curly brackets in (1) lies between 0 and 1. As a consequence, the inequality chain

$$0 \leq |\beta(x)| \leq |\beta_x| \quad (2)$$

holds true, providing readily interpretable bounds for the magnitude of the marginal effect $\beta(x)$. As already mentioned, links with the existing literature are present. In particular, with reference to the context considered here, results in [2] state that $\beta(x)$ and β_x must share the same sign, whereas those in [4] prove - in a setting where X is binary but that can be extended to the continuous case - that the magnitude of $\beta(x)$ is limited by that of the conditional effect β_x . Clearly, the combination of these two findings is equivalent to (2). However, as stated in the Introduction, the inequality chain in (2) is not optimal. Indeed, a higher value replacing 0 as the lower bound can in principle be found.

In order to identify such a lower bound, a more thorough study of the mathematical properties of the $\beta(x)$ function is necessary. To this end, it is convenient to note that (1) can be rewritten as

$$\beta(x) = \beta_x \left\{ 1 - \frac{q_1 \ell(x)}{q_2 \ell^2(x) + q_3 \ell(x) + q_4} \right\}, \quad (3)$$

where $\ell(x) = \exp(\beta_0 + \beta_x x)$ and

$$\begin{aligned} q_1 &= \exp(\gamma_0) \{ \exp(\beta_w) - 1 \}^2 \\ q_2 &= \exp(\beta_w) \{ \exp(\gamma_0) + 1 \} \{ \exp(\gamma_0) + \exp(\beta_w) \} \\ q_3 &= \exp(\beta_w) \{ 1 + \exp(\gamma_0) \}^2 + \{ \exp(\beta_w) + \exp(\gamma_0) \} \{ 1 + \exp(\beta_w) \exp(\gamma_0) \} \\ q_4 &= \{ 1 + \exp(\beta_w) \exp(\gamma_0) \} \{ 1 + \exp(\gamma_0) \}. \end{aligned}$$

The four coefficients above are always positive, with $q_3 \geq q_1$ confirming that the curly bracket term is in principle bounded between 0 and 1 like in (1). An advantage of Equation (3) is that it relates the marginal effect $\beta(x)$ directly to the parameters of the conditional logistic model without relying on Bayes' formula, which is needed to compute the probabilities in the $\Delta_w(x)$ term in (1).

Under this alternative formulation, it is easier to show (see the Appendix for the proof) that

$$\lim_{x \rightarrow \pm\infty} \beta(x) = \beta_x, \quad (4)$$

and that $\beta(x)$ has a single stationary point $\beta(x^*)$, where

$$x^* = \frac{1}{\beta_x} \left\{ \log \sqrt{\frac{1 + \exp(\beta_w) \exp(\gamma_0)}{\exp(\beta_w) \{ \exp(\beta_w) + \exp(\gamma_0) \}}} - \beta_0 \right\} \quad (5)$$

and γ_0 denotes the logistic transformation $\log(P(W = 1)/P(W = 0))$. Given the limits in (4) and the fact that $\beta(x)$ is a continuous function with a single stationary point, one can update (2) in the optimal way

$$|\beta(x^*)| \leq |\beta(x)| \leq |\beta_x|. \quad (6)$$

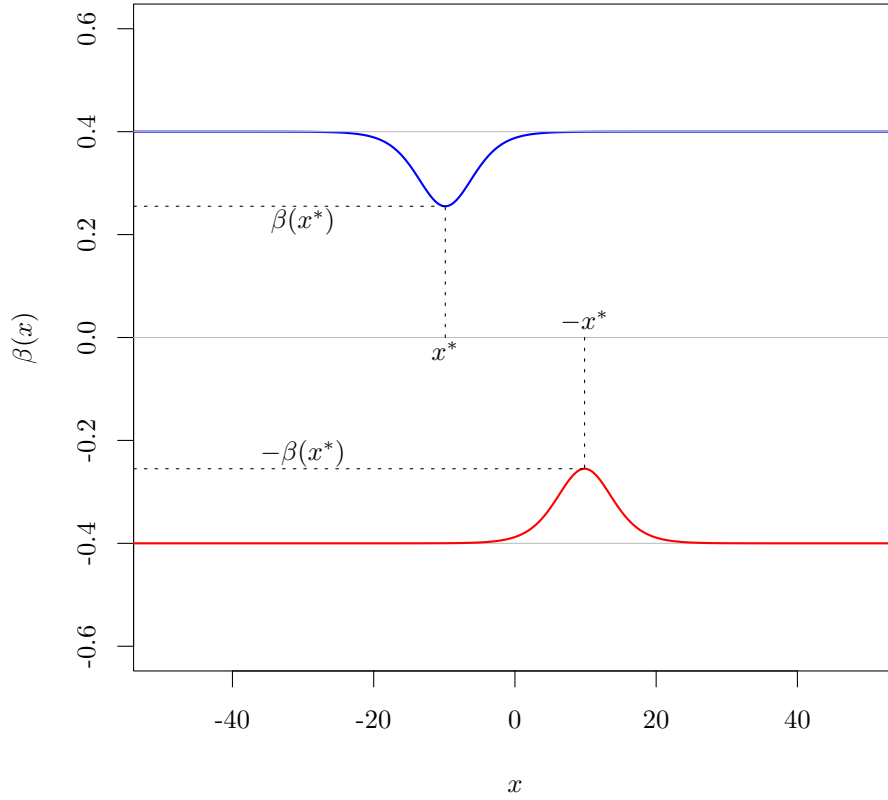


Fig. 1 Plot of the $\beta(x)$ (blue) and $\tilde{\beta}(x)$ (red) functions for $\gamma_0 = -1$, $\beta_0 = 2$, $\beta_w = 3$ and $\beta_x = \pm 0.4$.

It is straightforward to conclude that the stationary point is either a maximum if β_x is negative or a minimum if β_x is positive. An example is depicted in Figure 1, where two marginal effect functions are reported for two parameter sets differing in the sign of β_x only. Specifically, we have $\gamma_0 = -1$, $\beta_0 = 2$ and $\beta_w = 3$, with the blue curve referring to $\beta_x = 0.4$ and the red curve referring to $\tilde{\beta}_x = -0.4$. These curves are denoted by $\beta(x)$ and $\tilde{\beta}(x)$ respectively.

3 Symmetry of marginal effects

The $\beta(x)$ and $\tilde{\beta}(x)$ curves also present some interesting features related to symmetry. In particular, since the four coefficients q_1 , q_2 , q_3 and q_4 do not depend on β_x and $\ell(\beta_0 + \beta_x x) = \ell(\beta_0 + \tilde{\beta}_x(-x))$, by looking at the formulation in Equation (3) it is immediate to verify that

$$\tilde{\beta}(x) = -\beta(-x).$$

This pairwise symmetry holds in general and not only for the point x^* , for which an explicit mention is reported in Figure 1. Notice that the result above does not mean that reversing the conditional effect β_x will reverse the marginal effects computed at the *same* points (that would read $\tilde{\beta}(x) = -\beta(x)$). Instead, a change in the sign of β_x will reverse marginal effects computed at *opposite* points. This remark is important since marginal effect functions are not, in general, symmetrical with respect to the vertical axis traced in $x = 0$. Nevertheless, symmetry holds with respect to the axes corresponding to their stationary points. With reference to Figure 1, we have

$$\beta(x^* + x) = \beta(x^* - x)$$

and, analogously, $\tilde{\beta}(-x^* + x) = \tilde{\beta}(-x^* - x)$. This implies that marginal effects at points equidistant from the stationary point take the same value. A proof is contained in the final part of the Appendix.

4 Appendix

Under the formulation reported in Equation (3), the computation of the limits in (4) is immediate. Indeed, the ratio in the curly bracket term in (3) is a simple polynomial equation in $\ell(x)$ with

$$\lim_{x \rightarrow -\infty} \ell(x) = 0 \quad \lim_{x \rightarrow +\infty} \ell(x) = +\infty$$

if $\beta_x > 0$ and

$$\lim_{x \rightarrow -\infty} \ell(x) = +\infty \quad \lim_{x \rightarrow +\infty} \ell(x) = 0$$

if $\beta_x < 0$. Furthermore, simple algebraic developments show that

$$\frac{d}{dx} \beta(x) = -\beta_x^2 \ell(x) q_1 \left\{ \frac{-q_2 \ell^2(x) + q_4}{[q_2 \ell^2(x) + q_3 \ell(x) + q_4]^2} \right\}.$$

Equating to zero such a derivative is equivalent to solve

$$-q_2 \ell^2(x) + q_4 = 0,$$

whose only admissible solution in x is given by (5).

To prove the symmetry of $\beta(x)$ with respect to the vertical axis in $x = x^*$, it suffices to bear in mind that $\ell^2(x) = q_4/q_2$ and notice that $\ell(x^* + x) = \ell(x^*) \exp(\beta_x x)$. Therefore, it is possible to write

$$\beta(x^* + x) = \beta_x \left\{ 1 - \frac{q_1 \ell(x^*) \exp(\beta_x x)}{q_4 \{\exp(2\beta_x x) + 1\} + q_3 \ell(x^*) \exp(\beta_x x)} \right\}.$$

Similarly, since $\ell(x^* - x) = \ell(x^*) \exp(-\beta_x x)$, we have

$$\beta(x^* - x) = \beta_x \left\{ 1 - \frac{q_1 \ell(x^*) \exp(-\beta_x x)}{q_4 \{\exp(-2\beta_x x) + 1\} + q_3 \ell(x^*) \exp(-\beta_x x)} \right\}.$$

Multiplying by $\exp(2\beta_x x)$ both the numerator and the denominator of the ratio in the above equation returns the expression of $\beta(x^* + x)$, which proves symmetry. Analogous considerations holds for $\tilde{\beta}(x)$.

References

- [1] W. G. Cochran. The omission or addition of an independent variate in multiple linear regression. *Supplement to J. R. Statist. Soc.*, 5(2):171–176, 1938.
- [2] D. R. Cox and N. Wermuth. A general condition for avoiding effect reversal after marginalization. *J. R. Statist. Soc. B*, 65(4):937–941, 2003.
- [3] D. Y. Lin, B. M. Psaty, and R. A. Kronmal. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, 54(3):948–963, 1998.
- [4] J. M. Neuhaus and N. P. Jewell. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, 80(4):807–815, 1993.
- [5] A. Sjölander, E. Dahlqvist, and J. Zetterqvist. A note on the noncollapsibility of rate differences and rate ratios. *Epidemiology*, 27(3):356–359, 2016.
- [6] E. Stanghellini and M. Doretti. On marginal and conditional parameters in logistic regression models. *Biometrika (forthcoming)*, 2019.

DNA mixtures: a case study involving a Romani reference population

Mixture di DNA: un caso di studio riguardante una popolazione di riferimento dei Rom

Francesco Dotto, Julia Mortera and Vincenzo Pascali

Abstract Here we present an Italian criminal case to show how statistical methods can be used to extract information from a series of mixed DNA profiles. The case involves several different individuals and a set of different DNA traces. First, a brief description of the case is provided. Secondly, we introduce some heuristic tools that can be used to evaluate the data and briefly outline the statistical model used for analysing DNA mixtures. Finally, we illustrate some of the findings on the case and discuss further directions of research.

Abstract *L'obiettivo del presente contributo è quello di mostrare come, tramite l'utilizzo di opportuni metodi statistici, sia possibile estrarre informazioni da un insieme di misture di profili di DNA. A tal fine, verrà utilizzato, come caso di studio un caso di cronaca italiana. In particolare verrà, come prima cosa, fornita una breve descrizione del caso. In secondo luogo verranno introdotti alcuni strumenti euristici utili all'analisi dei dati e verrà fornita una descrizione del modello statistico utilizzato. Infine verranno illustrati e commentati i risultati e i possibili sviluppi di ricerca futuri.*

Key words: Bayesian networks, combining evidence, DNA mixtures, forensic statistics, likelihood ratio

Francesco Dotto

Dipartimento di Economia, Università degli Studi di Roma Tre, Via Silvio D'Amico, 77, 00145 Roma. e-mail: francesco.dotto@uniroma3.it

Julia Mortera

Università degli Studi di Roma Tre, Via Silvio D'Amico, 77, 00145 Roma. e-mail: julia.mortera@uniroma3.it

Vincenzo Pascali

Istituto di Sanità Pubblica, Università Cattolica del Sacro Cuore, Largo Francesco Vito, 1, 00168 Roma. e-mail: vincenzolorenzo.pascali@unicatt.it

1 Introduction

1.1 Case history

In a small village in North Italy, four men broke into a private courtyard in an attempt to commit a theft. They were noticed by two bystanders and fled. The bystanders alerted the local police station. To escape, the four rogues stopped a car driver, hijacked his car and then disappeared. The next day, the car was found, concealed in a country road, and a baseball cap was retrieved from the vehicle's seat. The cap did not apparently belong to the car owner. The cap was brought to the local DNA laboratory, inspected under UV light and seven fabric samples were excised from its inner side. We will denote these samples B_1, B_2, \dots, B_7 . Five individuals - two of them were of Romani ethnicity, together with the car owner - were subsequently investigated. We refer to these persons of interest (PoI). A saliva swab was taken from all PoI to obtain their DNA profiles. These profiles were compared to the DNA evidence from the baseball cap.

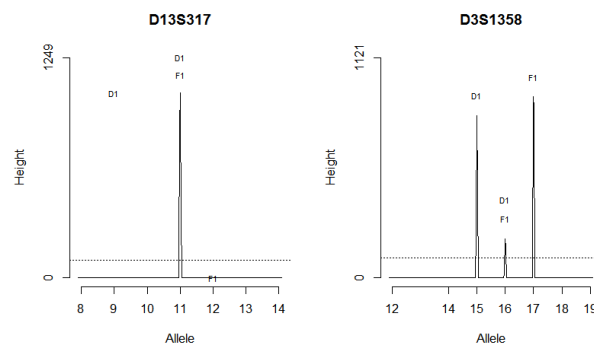
1.2 Genetic background

A DNA molecule consists of two strands consisting of *nucleobases* (bases) adenine (A), thymine (T), cytosine (C), or guanine (G). DNA is usually represented by a sequence of bases, “words” in the DNA alphabet A, T, C and G. Human DNA in cell nuclei has 23 pairs of chromosomes, 22 autosomal and two sex-determining chromosomes. An area on a chromosome is called a *locus*. The DNA composition *i.e.* a particular sequence of the four bases at a given locus is called an *allele*. A genotype of an individual at a locus is an unordered pair of alleles. Rather than identifying a person through their entire DNA sequence, for forensic purposes a list of genotypes at 15–20 chosen *markers* at selected *loci*, are considered. In particular, current technology uses *short tandem repeat* (STR) markers, which are markers with alleles generally given by integers. So, if an STR allele takes value 5, then a certain word in the DNA alphabet, *e.g.* AGAT, is repeated 5 times at that locus. If there is a partial repeat then the size in bases of the partial repeat is given after a decimal point. For example, the allele named 9.3 consists of 9 repeats and a partial repeat of 3 base pairs. In statistical terms, a locus corresponds to a random variable while an allele is its realised state. Additionally, for forensic markers, we can assume independence of alleles within and across markers, as they are located on different chromosomes. For further details see [2], [3] and [6] and references therein.

1.3 DNA mixtures

In the case analysis examined here, the evidence does not concern a simple single source DNA profile, but several complex *DNA mixture* samples. DNA mixture samples contain DNA from multiple contributors, as *e.g.* the $B1, \dots, B7$ traces found on the baseball cap. The questions posed to a forensic statistician could be: “Given a DNA mixture, who are the contributors?” or “Is a suspect a contributor to one or more DNA mixtures?”. To answer these type of queries, an approach is to jointly model the DNA profiles of the potential known and unknown contributors to the mixtures and the peak heights in electropherograms (EPG). The EPGs are produced from the amplification –using the polymerase chain reaction, PCR– of the DNA in each sample. The EPG is a chart which displays, for each marker, the alleles detected and the corresponding peak heights. It is based on capillary electrophoresis and measures the responses in *relative fluorescence units* (RFU). The peak height is a measure of the amount of the allele in the amplified sample expressed in RFUs. The peak heights in the EPG contain important information about the composition of the DNA mixture. For further details on DNA profiling the reader is referred to [2] and [10]. Figure 1 shows a pictorial representation of the EPG for sample $B3$ at markers D3S1358 and D13S317. The labels D1 and F1 denote the genotypes of the two suspects. Marker D13S317 in sample $B3$ has a single peak at allele 11. Suspects D1 and F1 have genotypes (9,11) and (11,12), respectively, so only their allele 11 is amplified in sample $B3$. So, if D1 and F1 were contributors to $B3$, one of their alleles must not have been amplified, an artefact called a dropout. Whereas, for marker D3S1358, the EPG has peaks at alleles 15, 16, and 17. F1 has genotype (16,17) and D1 has genotype (15,16), so all their alleles are present for this marker. However, under the hypothesis that both D1 and F1 were contributors, it is highly unlikely that the EPG would yield such a small peak at allele 16, as D1 and F1 would each contribute a proportion of DNA to the peak height at 16.

Fig. 1 A pictorial representation of the electropherogram, showing the peak heights and corresponding alleles, for sample $B3$ at markers D13S317 and D3S1358. The genotypes of individuals D1 and F1 at these markers are also given.



2 A statistical model

2.1 The weight of evidence

Here we examine the evidence E consisting of the peak heights and alleles in the DNA mixture samples $B3$ and $B6$, together with the genotypes of the known individuals $D1$ and $F1$, *i.e.* $E = \{D1, F1, \text{DNA mixture}\}$. We wish to consider how the evidence E affects the comparison between the prosecution $H_p : D1 \& U1 \& U2$ and defence hypotheses $H_d : U1 \& U2 \& U3$ where, H_p claims that $D1$ and two unknown individuals, U_1 and U_2 contributed to the DNA mixture, whereas, H_d states that 3 distinct unknown individuals U_1 , U_2 and U_3 contributed to the mixture. The DNA profiles of the known individuals are considered fixed, whereas the DNA profiles of the unknown contributors are considered mutually independent and sampled from a suitable reference population. Similarly, a prosecution and defence hypothesis can be formulated for the second suspect $F1$, by substituting $F1$ for $D1$, above. The strength of the evidence is reported as a likelihood ratio ([1], [4] and [8])

$$\text{LR} = \frac{L(H_p)}{L(H_d)} = \frac{P(E|H_p)}{P(E|H_d)} \quad (1)$$

or the weight of evidence: $\text{WoE} = \log_{10}(\text{LR})$.

2.2 The Gamma Model

Computation of the likelihood ratio (1) is based on the model described in [3]. The model assumes that the variability in the peak heights at an allele is independent of the variability at the other allelic positions when the model parameters and genotypes are considered fully known. The model takes into account the artefacts that can occur in the PCR amplification process: stutter, whereby a proportion of a peak belonging to allele a appears as a peak at allele $a - 1$; and dropout, when alleles are not observed because the peak height is below a detection threshold C . A further artefact, termed dropin (which refers to the occurrence of small unexpected peaks due to contamination) is captured by including extra unknown contributors. Now, consider allele a , the variability of the peak height Z_a at a can be expressed as the gamma distribution

$$Z_a \sim \Gamma \left\{ \frac{1}{\sigma^2} D_a(\phi, \xi, \mathbf{n}), \mu \sigma^2 \right\}, \quad (2)$$

where ϕ_i denotes the *proportion* of DNA originating from individual i prior to PCR amplification, n_{ia} is the number of type a alleles for individual i , and $D_a(\phi, \xi, \mathbf{n}) = (1 - \xi) \sum_i \phi_i n_{ia} + \xi \sum_i \phi_i n_{i,a+1}$ are the effective allele counts after stutter. In a sample from a single donor, where no dropout or stutter has occurred, μ is the mean peak height and σ is the coefficient of variation. For example, $\sigma = 0.58$ corre-

sponds to the standard deviation of the peak being 58% of its mean μ . The back-stutter parameter ξ determines the mean proportion of stutter that may be observed in the allelic position one repeat less. Here ξ is the ratio of the stutter peak with respect to the parent plus the stutter peak, rather than the more commonly used ratio between the stutter peak and the parent peak [9]. The evidence E consists of the peak heights \mathbf{z} as observed in the EPGs, as well as any potential genotypes of known individuals. For given genotypes of the contributors, expressed as allele counts $\mathbf{n} = (n_{ia}, i = 1, \dots, I; a = 1, \dots, A)$, given proportions ϕ , and given values of the parameters $\psi = (\mu, \xi, \sigma)$, all observed peak heights are independent and for a given hypothesis H , the full likelihood is obtained by summing over all possible combinations of genotypes \mathbf{n} with probabilities $P(\mathbf{n}|H)$ associated with H :

$$L(H) = \Pr(E|H) = \sum_{\mathbf{n}} L(\psi|\mathbf{z}, \mathbf{n}) P(\mathbf{n}|H) \quad (3)$$

where

$$L(\psi|\mathbf{z}, \mathbf{n}) = \prod_m \prod_a L_{ma}(z_{ma}) \quad (4)$$

and

$$L_{ma}(z_{ma}) = \begin{cases} g\{z_{ma}; \sigma^{-2} D_a(\phi, \xi, \mathbf{n}), \mu \sigma^2\} & \text{if } z_{ma} \geq C \\ G\{C; \sigma^{-2} D_a(\phi, \xi, \mathbf{n}), \mu \sigma^2\} & \text{otherwise,} \end{cases} \quad (5)$$

with g and G denoting the gamma density and cumulative distribution function respectively. The number of terms in the sum (3) is huge for a hypothesis which involves several unknown contributors to the mixture. However, it can be calculated efficiently by Bayesian network techniques that represent the genotypes using a Markovian structure, the allele counts for each individual being modelled sequentially over the alleles. The maximum likelihood estimate (MLE) parameters are obtained using the R package `DNAmixtures` [5] which interfaces to the HUGIN API (Hugin Expert A/S, 2012) through the Rpackage `RHugin` [7].

3 Results

This case concerns PoI from the Romani and Italian population, so we used the following population allele frequency databases: Macedonian Romani (M); Portuguese Romani (PO); Eastern Slovakian Romani (ES) and the Italian Caucasian (IT). Table 1 shows the parameter estimates and the likelihood ratio for comparing the hypotheses $H_p : D1 \& U1 \& U2$ vs. $H_d : U1 \& U2 \& U3$ and $H_p : F1 \& U1 \& U2$ vs. $H_d : U1 \& U2 \& U3$ for the DNA mixtures in samples *B3* and *B6*.

Note that the estimates for μ and σ do not, as expected, vary among the different reference populations and the different hypotheses. Furthermore, the stutter parameter ξ is almost null, indicating that potential stutter had been filtered out of the data. This can be dangerous as a true peak might be confused with a stutter peak. The proportion of DNA contributed by the main suspect *F1* is small for sample *B3*,

Table 1 Estimated parameters and likelihood ratios for the prosecution and defence hypothesis for the DNA mixtures in samples *B3* and *B6*.

Sample	Ref. Pop.	$H_p : F1 \& U1 \& U2$ vs. $H_d : U1 \& U2 \& U3$							$H_p : D1 \& U1 \& U2$ vs. $H_d : U1 \& U2 \& U3$						
		μ	σ	ξ	ϕ_{U1}	ϕ_{U2}	ϕ_{F1}	LR	μ	σ	ξ	ϕ_{U1}	ϕ_{U2}	ϕ_{D1}	LR
B3	IT	536	0.91	0.00	0.43	0.42	0.15	1.73	537	0.90	0.00	0.54	0.22	0.24	22.23
B3	M	535	0.88	0.00	0.45	0.45	0.10	1.25	538	0.89	0.00	0.39	0.39	0.22	10.08
B3	PO	534	0.88	0.00	0.45	0.45	0.11	1.43	535	0.88	0.00	0.42	0.42	0.17	3.10
B3	ES	536	0.90	0.00	0.42	0.42	0.15	1.75	537	0.90	0.00	0.49	0.29	0.22	18.61
B6	IT	542	0.60	0.00	0.61	0.38	0.01	1.01	551	0.68	0.00	0.47	0.25	0.28	45.16
B6	M	541	0.61	0.00	0.62	0.38	0.00	1.00	551	0.70	0.00	0.41	0.31	0.28	30.43
B6	PO	540	0.60	0.00	0.57	0.43	0.00	1.00	550	0.68	0.01	0.38	0.38	0.25	9.20
B6	ES	554	0.66	0.10	0.64	0.26	0.11	1.37	556	0.76	0.10	0.33	0.33	0.34	73.23

$\phi_{F1} \leq 0.15$, so there is very little evidence that *F1* is a contributor to the two DNA traces. The weight of evidence against *F1* is also small, $LR \leq 1.75$ and should certainly not lead to his incrimination. Whereas, the proportion of DNA contributed by the PoI, *D1*, is slightly larger $\phi_{D1} \in [0.17 - 0.34]$ and the weight of evidence against *D1* is quite substantial in both *B3* and *B6*, for all databases except the Portuguese (3.1). For sample *B6*, using the Eastern Slovakian (ES) database yields a large weight of evidence $WoE=73.23$ in favour of *D1* being among the contributors. This illustrates how the weight of evidence can vary when using different reference populations, some leading to exonerating and some to condemning a suspect. This issue will be explored further in future work.

References

1. Balding, D.J., Steele, C.D.: Weight-of-evidence for forensic DNA profiles. John Wiley & Sons (2015)
2. Butler, J.M.: Typing: Biology, technology, and genetics of STR markers. New York, ed 2 (2005)
3. Cowell, R.G., Graversen, T., Lauritzen, S.L., Mortera, J.: Analysis of forensic DNA mixtures with artefacts (with discussion). JRSS: Series C (Applied Statistics) **64**(1), 1–48 (2015)
4. Good, I.J.: Probability and the weighing of evidence (1950)
5. Graversen, T.: DNAmixtures: Statistical inference for mixed traces of dna. R package version 0.1-0, dnadmixture. r-forge. r-project. org (2013)
6. Graversen, T.: Statistical and computational methodology for the analysis of forensic DNA mixtures with artefacts. Ph.D. thesis, Oxford University, UK (2014)
7. Konis, K.: RHugin. R package version 7.8 (2014)
8. Lindley, D.: A problem in forensic science. Biometrika **64**(2), 207–213 (1977)
9. Tvedebrink, T., Eriksen, P.S., Asplund, M., Mogensen, H.S., Morling, N.: Allelic drop-out probabilities estimated by logistic regression—further considerations and practical implementation. Forensic Science International: Genetics **6**(2), 263–267 (2012)
10. Tvedebrink, T., Eriksen, P.S., Mogensen, H.S., Morling, N.: Evaluating the weight of evidence by using quantitative short tandem repeat data in DNA mixtures. JRSS: Series C (Applied Statistics) **59**(5), 855–874 (2010)

Pivotal seeding for K -means based on clustering ensembles

Inizializzazione pivotale dell'algoritmo delle K -medie tramite raggruppamento con metodi di insieme

Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli

Abstract Despite its large use, one major limitation of K -means algorithm is the impact of the initial seeding on the final partition. We propose a modified version, using the information contained in a co-association matrix obtained from clustering ensembles; such matrix is given as input for a set of pivotal methods, implemented in the `pivmet` R package, used to perform a pivot-based initialization step. Preliminary results concerning the comparison with the classical approach and other clustering methods are discussed.

Abstract Nonostante l'ampio uso che ne viene fatto, una delle maggiori limitazioni dell'algoritmo delle K -medie riguarda l'impatto della scelta dei semi iniziali sulla partizione finale. Viene proposta una variante che usa l'informazione contenuta in una matrice di co-associazione ottenuta per mezzo dei cosiddetti metodi di insieme: tale matrice viene utilizzata da alcuni metodi pivotali, implementati nel pacchetto R `pivmet`, impiegati per definire una fase iniziale basata su unità pivotali. Vengono discussi alcuni risultati preliminari riguardanti il confronto con l'approccio classico e con altri metodi di raggruppamento.

Key words: Clustering, pivotal method, seeding

1 Introduction

Data clustering is a complex task in the framework of unsupervised learning. It deals with finding a structure in a collection of unlabelled data, and specifically, it aims at identifying groups of objects that are similar according to a suitable proximity measure [6]. Among the partitional methods, the K -means algorithm, which aims

Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche, 'Bruno de Finetti',
Università degli Studi di Trieste, Via Tigor 22, 34124 Trieste, Italy, e-mail: legidi@units.it, rpap-
pada@units.it, francesco.pauli@deams.units.it, nicola.torelli@deams.units.it

at minimizing the overall distance of the n data points from their respective cluster centers, is the most popular and the simplest one [9]. As a matter of fact, each algorithm has its own strengths and weaknesses, and yet no single algorithm is able to outperform all the others in terms of the quality of clustering results, which are often driven by subjective choices related to the problem at hand. In the case of K -means clustering, for instance, several limitations arise when dealing with patterns characterized by non-spherical shapes, different sizes or density. The method is also significantly sensitive to the initial randomly selected cluster centers. The K -means algorithm is usually run multiple times to reduce this effect.

Recent works explore cluster ensemble methods in order to address the problem of combining multiple clusterings of the same set of objects into a single final clustering solution (see, among the others, [10, 8]). Although cluster ensembles have been shown to be beneficial in a wide variety of scenarios, the issue of combining different clustering results is non-trivial and poses new challenges. First, there are several possible ways to perform the so-called evidence accumulation-based clustering, depending on the mechanism for generating partitions, the parameters involved, and the criterion used to obtain the final solution or *consensus* partition, which is then expected to improve the quality of results as compared to a single clustering approach. A possible way to summarize the information coming from multiple techniques is to obtain a *co-association* matrix (see e.g. [7]), taking the co-occurrences of pairs of points in the same cluster across distinct solutions, thus leading to a new measure of similarity between the objects; such matrix is then used to deliver the final consensus clustering. Second, it is usually difficult to establish whether, and to which extent, the final consensus clustering is ‘better’ than those which were combined.

Here, we discuss how to use the co-association matrix to find some specific units, called *pivots*, which can be representative for the group they belong to, for a given partition of the data. The algorithm of Maxima Units Search (MUS), introduced in [4] and further developed in [3], has been shown to give satisfactory performance for a small number of clusters.

In Section 2 we discuss the issue of pivot identification and some alternative criteria are proposed. In Section 3 a modified version of classical K -means is presented, which uses pivotal units for initial seeding. A preliminary simulation experiment is illustrated in Section 4. Section 5 concludes.

2 Pivotal units based on clustering ensembles

Let $\mathbf{Y} = (y_1, \dots, y_n)$ be a set of n observations, where y_i may be defined over some d -dimensional space, $y_i \in \mathbb{R}^d$. Let \mathcal{P} represent a set of H partitions, which defines a *clustering ensemble*, $\mathcal{P} = \{P^1, P^2, \dots, P^H\}$, where each $P^h = \{C_1^h, C_2^h, \dots, C_{k_h}^h\}$ is a partition of the data set \mathbf{Y} into k_h disjoint clusters. In general, the value of k_h for different clustering runs can be either the same or different. In our study we only consider the former case, i.e. $K = k_h$, for all $h \in \{1, \dots, H\}$. As mentioned before,

a common strategy to combine the H clustering solutions is to map the partitions in \mathcal{P} into a co-association matrix, whose generic element is computed as

$$c_{i,j} = \frac{1}{H} \sum_{h=1}^H |P^h(y_i) = P^h(y_j)|,$$

where $|\cdot|$ denotes the indicator function, and $P^h(y_i)$ represents the associated label of the object y_i in the partition P^h . Thus, the value in each position (i, j) of this matrix expresses the fraction of times the two data points y_i and y_j are classified in the same cluster across all partitions in \mathcal{P} . The matrix $C = (c_{i,j})$ fulfils the conditions of a similarity matrix and it is expected to contain a non-negligible number of zeros.

Given a large and sparse matrix, MUS algorithm [3] seeks those elements, among a pre-specified number of candidate pivots, whose corresponding rows contain more zeros compared to all other units. In this framework, we exploit the fact that the entries in C corresponding to units that are very ‘dissimilar’ are likely to be zero.

Now, let C_1, \dots, C_K be a reference partition of the data obtained, for instance, by applying an agglomerative hierarchical algorithm and taking the solution with K groups. (The sensitivity of the proposed method to the choice of this reference partition will be the subject of future research.) Starting from C_1, \dots, C_K , the MUS procedure is used to find K pivots, $y_{i_1}, \dots, y_{i_K} \in \mathbf{Y}$, one for each group, which exhibit the highest degree of separation. Ideally, the submatrix of C with only the rows (columns) of the selected pivotal units is identical or nearly identical. It is worth noting that pivots identification is not an easy task, and MUS algorithm can be computationally demanding for large n and when K increases ($K > 4$). An alternative strategy consists of choosing a pivotal unit from each cluster in such a way it is as far as possible from units that might belong to the other groups and/or as close as possible to units that belong to the same group, according to a suitable objective function. The pivot y_{i_k} for group C_k , $k \in \{1, \dots, K\}$ can be chosen so that it maximizes one of the quantities:

$$(i) \sum_{j \in C_k} c_{i_k, j}; \quad (ii) \sum_{j \in C_k} c_{i_k, j} - \sum_{j \notin C_k} c_{i_k, j}, \quad (1)$$

where (i) and (ii) express the global within similarity the discrepancy between global within and between similarities, respectively.

3 Seeding in k -means clustering using pivotal units

In [5] a co-association matrix C is obtained by random initializations of the K -means algorithm (with K fixed) leading to variations in the final configurations. A new partition is obtained by setting as initial centers the pivots based on C and selected via the MUS algorithm. Such approach has been proved to be advantageous in case of clusters of unbalanced sizes and non-spherical shapes (for an alternative approach to careful seeding see [1]).

In this section, we extend the procedure summarized above, namely the MUSK-means algorithm, in order to include the alternative pivotal selection criteria described in the previous section. Given n data points, (y_1, \dots, y_n) , the number of groups K and the size of clustering ensemble H , the `piv_KMeans` procedure consists of the following steps.

Algorithm: `piv_KMeans`

1. Initialization

- a. Generate the clustering ensemble \mathcal{P} of H partitions, where each clustering is the result of a K -means run with randomly selected cluster centers μ_1, \dots, μ_K .
- b. Build the co-association matrix $C = (c_{i,j})$, $i, j = 1, \dots, n$, using \mathcal{P} .
- c. Fix an arbitrary partition of the data set in K groups, G_1, \dots, G_K , and apply a pivotal method to C choosing among MUS, (i), (ii), to find the pivots y_{i_1}, \dots, y_{i_K} ; then, set $\mu_k = y_{i_k}$, for all $k = 1, \dots, K$.

2. Update

Iterate until the cluster assignments stop changing:

- a. Assign each y_i to the cluster G_k whose centre μ_k is closest
- b. Recompute cluster centers as the centroid of the observations in each group.

Note that the proposed algorithm differs from classical K -means in the initialization step, where the initial random cluster assignment is replaced by steps 1a., 1b., 1c., involving the generation of cluster ensembles and the selection of a set of pivotal units.

4 Simulation experiment

A simulation experiment is carried out to explore the performance of our method via the `piv_KMeans` function listed in the `pivmet` R package [2]. This function is similar to the standard `kmeans` function but allows for some optional arguments related to the pivotal criteria we may choose.

As already mentioned, one of the drawbacks of K -means is its inefficiency in distinguishing between groups of unbalanced sizes. For illustration purposes, we simulate (a) three bivariate Gaussian distributions with 20, 100 and 500 observations, respectively (b) ‘two-sticks’ shaped groups of 30 and 370 observations, respectively (Fig. 1). The plots with titles ‘`piv_KMeans`’ refer to the pivotal criteria MUS, (i) or `maxsumint`, (ii) or `maxsumdiff`, where the labels 1, 2, and 4 follow the order used in the R function; moreover, we consider Partitioning Around Medoids (PAM) method via the `pam` function of the `cluster` package and agglomerative hierarchical clustering (`agnes`), with average, single, and complete linkage. The partitions from the classical K -means are obtained using multiple random seeds. Group centers and pivots are marked via asterisks and triangles symbols, respectively.

In the first scenario the classical K -means and the PAM tend to split the cluster with the highest density in two separate clusters, whereas the partition identified by

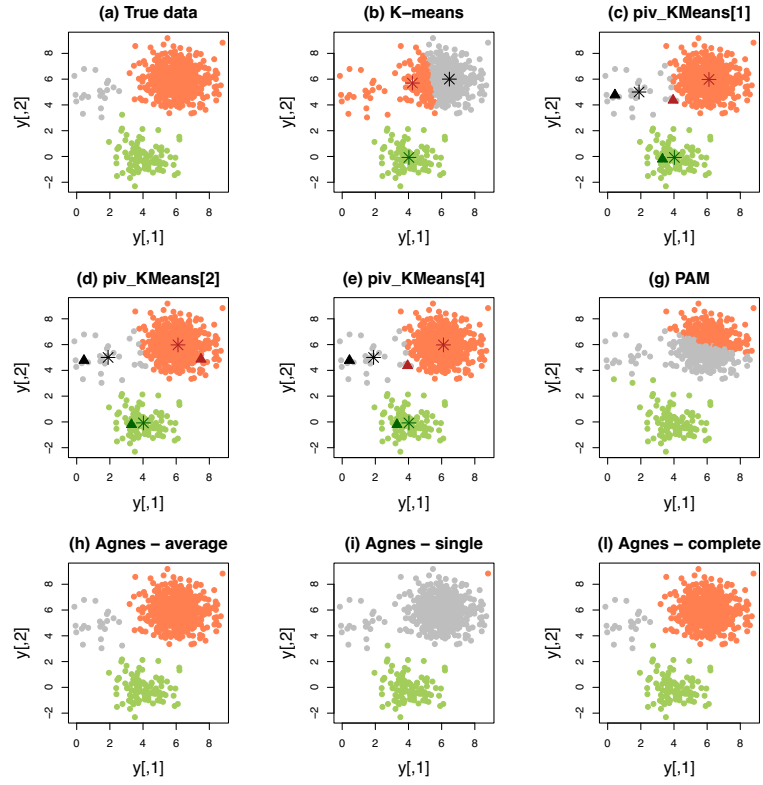
the agglomerative hierarchical clustering (single-linkage) is far from the true one; conversely, the cluster compositions identified by `piv_KMeans` with three pivotal criteria and by agglomerative clustering (average and complete-linkage) show a greater agreement with the true partition. For `piv_KMeans` the final centers are close to the pivotal units used for the seeding. The same considerations remain almost valid for the second scenario, but here the solution from single-linkage is quite good, whereas the solution obtained from complete-linkage is similar to those given by classical K -means and PAM. The Adjusted Rand Index (ARI) is computed for 1000 iterations as a measure of agreement between the resulting clustering and the true data partition. Among the considered criteria, `piv_KMeans` outperforms the classical algorithm and the PAM method in both scenarios; in particular, `piv_KMeans[1]` (MUS method) and `piv_KMeans[2]` (Eq. (1)-(i)) yield the highest ARI.

5 Conclusions

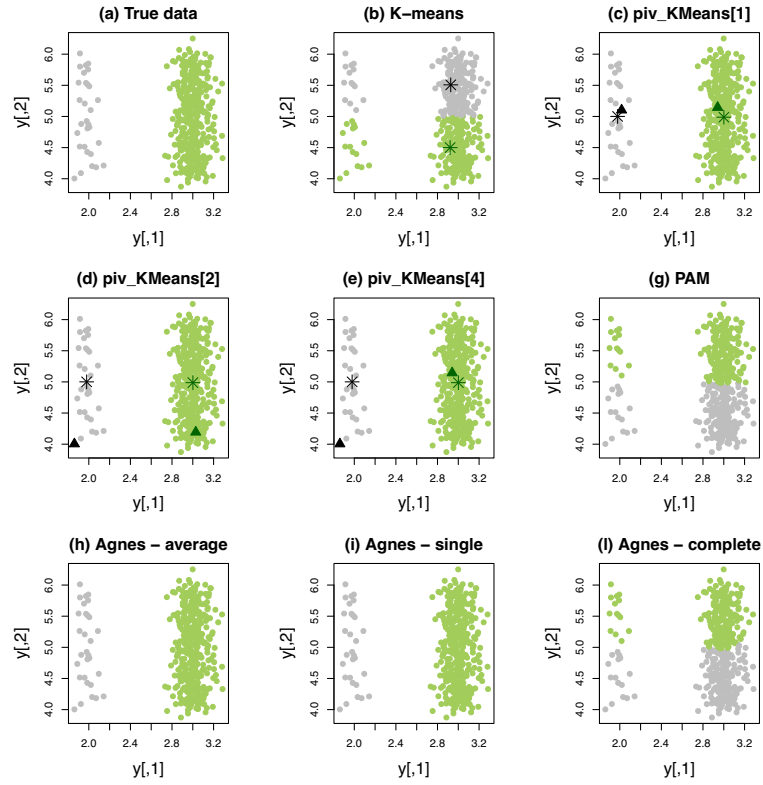
We propose a modified K -means algorithm exploiting a pivotal-based initialization step, which can be applied to reduce the impact of random seeding. A simulation experiment reveals that our procedure outperforms the classical approach in terms of the quality of the clustering solutions in the considered scenarios. We also evaluate and compare different methods for identifying pivotal units based on clustering ensembles.

References

1. Arthur, D., Vassilvitskii, S.: `k-means++`: The advantages of careful seeding. In: Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms, pp. 1027-1035 (2007)
2. Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: `pivmet`: Pivotal methods for Bayesian relabelling and k -means clustering. R package version 0.1.1. Available on CRAN (2018)
3. Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: Maxima Units Search (MUS) algorithm: methodology and applications. In: Perna, C., Pratesi, M., Ruiz-Gazen A. (eds.) *Studies in Theoretical and Applied Statistics*, pp. 71–81 (2018)
4. Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: Relabelling in Bayesian mixture models by pivotal units. *Stat. Comput.* **28**(4), 957–969 (2018)
5. Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: K -means seeding via MUS algorithm. In: Abbruzzo, A., Brentari, E., Chiodi, M., and D. Piacentino (eds.), *Book of Short Papers SIS 2018*, 7 pp. ISBN:9788891910233.
6. Everitt, B. S.: *Cluster Analysis*. Heinemann, London (1981)
7. Fred, A. L., Jain, A. K.: Combining multiple clusterings using evidence accumulation. *IEEE Trans Pattern Anal Mach Intell* **27**(6), 835–850 (2005)
8. Ghosh, J., Acharya, A.: *Cluster Ensembles: Theory and Applications*. In: Aggarwal C. C. and Reddy C. K. (eds.) *Data Clustering: Algorithms and Applications*, pp. 551–570. CRC Press, Boca Raton FL (2014)
9. Jain, A.: Data clustering: 50 years beyond K -means. *Patt. Recog. Lett.* **31**(8), 651–666 (2010)
10. Strehl, A., Joydeep G.X: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583–617 (2002)



(a) Bivariate Gaussian



(b) 'Sticks' data

Fig. 1: Clustering solutions from scenarios (a) and (b)– See the text.

Optimal scoring of partially ordered data, with an application to the ranking of smart cities

Scoring ottimale di dati parzialmente ordinati, con un'applicazione al ranking delle smart city

Marco Fattore, Alberto Arcagni, Filomena Maggino

Abstract In this paper, we propose a solution to the problem of scoring and ranking partially ordered data, by exploiting the spectral properties of so-called matrices of *mutual ranking probabilities*, a class of matrices which comprise and convey information on the dominance among statistical units. The procedure is optimal in many respects and overcomes the limitations of other ranking tools, which may fail to deliver acceptable solutions, even in trivial cases. We show the algorithm in action, on real data pertaining to the smartness of a subset of European cities.

Abstract In questo articolo, viene proposto un algoritmo per l'attribuzione di score a dati parzialmente ordinati e per l'estrazione di ranking, basato sull'analisi delle proprietà spettrali delle cosiddette matrici di *mutual ranking probability*. La procedura è ottimale sotto diversi punti di vista e supera i limiti degli algoritmi attualmente disponibili, che possono generare ranking sub-ottimali, anche in casi molto semplici. Infine, si propone una breve applicazione a dati reali, relativi alla “smartness” di alcune città europee.

Key words: Multi-indicator system, Mutual ranking probability, Partially ordered set, Ranking, Scoring, Smart City.

Marco Fattore

Department of Statistics and Quantitative Methods, University of Milano-Bicocca, e-mail: marco.fattore@unimib.it

Alberto Arcagni

Department of Statistics and Quantitative Methods, University of Milano-Bicocca, e-mail: alberto.arcagni@unimib.it

Filomena Maggino

Department of Statistical Sciences, University of Roma - La Sapienza, e-mail: filomena.maggino@uniroma1.it

1 Introduction

In the practice of socio-economic statistics, it is increasingly frequent to deal with so-called ordinal *multi-indicator systems* (MISes), i.e. with complex sets of several attributes of an ordinal kind. Consider for example MISes provided by official surveys on well-being, quality-of-life and deprivation, or the problem of assessing the quality of services, based on non-numerical features and customers' satisfaction degrees. In general, the main (although, not necessarily the unique) goal is to turn the input data system into a ranking of the statistical units. While in the case of cardinal variables this can be often achieved quite effectively, by using various dimensionality reduction tools from linear and non-linear data analysis, in the ordinal setting just a few, and in many respects non-satisfactory, algorithms are currently available. In this short paper, we fill the gap and develop an optimal scoring and ranking algorithm, by integrating classical results from linear algebra (namely, the Singular Value Decomposition of non-negative matrices) and Partial Order Theory, which can be considered the “grammar” of ordinal data.

2 Partial orders and mutual ranking probabilities

A partially ordered set (or a *poset*) $\pi = (X, \leq)$ is a set X equipped with a partial order relation \leq , i.e. with a binary relation satisfying the properties of *reflexivity*, *antisymmetry* and *transitivity* [4, 11]. If $x \leq y$ or $y \leq x$, then x and y are called *comparable*, otherwise they are called *incomparable* (written $x \parallel y$). A partial order π where any two elements are comparable is called a *linear order* or a *complete order* or a *chain*. On the contrary, if any two elements of π are incomparable, then π is called an *antichain*. Given two partially ordered sets $\pi = (X, \leq_\pi)$ and $\sigma = (X, \leq_\sigma)$ on the same set X , we say that σ is an extension of π , if $y \leq_\pi x$ in π implies $y \leq_\sigma x$ in σ . In other terms, σ is an extension of π if it may be obtained from the latter by turning some incomparabilities into comparabilities. If σ is an extension of π and, in addition, it is a linear order, then it is called a *linear extension* of π . It can be proven that the set of all the linear extensions of a finite poset uniquely identifies the poset itself. Given two elements x_i and x_j of π , the so-called *mutual ranking probability* (MRP) p_{ij} between x_i and x_j is defined as the fraction of linear extensions where x_i is ordered below x_j [5, 6, 7]; one can then associate to the n -element poset π the $n \times n$ MRP matrix P , whose elements are defined as $P_{ij} = p_{ij}$. Notice that $P_{ij} = 1$ if and only if $x_i \leq x_j$, in the poset; this proves that the MRP determines the partial order relation.

3 The scoring and ranking algorithm

Building a ranking out of a poset π means defining an order-preserving map $r(\cdot) : \pi \rightarrow \lambda$, where λ is a target linear order. To achieve this, we go through an intermediate *score* function $s : \pi \rightarrow \mathbb{R}^+ \cup \{0\}$, which assigns, to each element of the input poset, a non-negative score that, informally speaking, should quantify how much it “dominates” the other elements. Notice that, in general, such a target linear order may comprise less elements than π ; in fact, if the nodes corresponding to two poset elements are *equivalent* in π (written $x \sim y$)¹, then their scores must be the same, for symmetry reasons. In order to get a consistent score system, and a consistent ranking as well, function $s(\cdot)$ must then satisfy two requirements:

1. it must be *strictly* order-preserving, i.e. $x \leq y \Rightarrow s(x) < s(y)$.
2. it must be “equivalence” preserving, i.e. $x \sim y \Rightarrow s(x) = s(y)$.

These two properties imply that: (i) if the input poset is a linear order λ , then the score function is injective and so the final ranking coincides with λ and (ii) if the input poset is an antichain α , then function $s(\cdot)$ is a constant, i.e. it collapses all of poset elements into the same score and the final ranking reduces to just a singleton. The main result of the paper is that a consistent scoring function is obtained by considering the first right singular vector of the MRP matrix associated to the input poset. More formally, the following result holds.

Proposition. Let π be a finite poset, with elements labeled as x_1, \dots, x_n , P be its MRP matrix, $\mathbf{v} = (v_1, \dots, v_n)^T$ be the first column of matrix V (first right singular vector) and $\mathbf{u} = (u_1, \dots, u_n)^T$ be the first column of matrix U (both \mathbf{v} and \mathbf{u} are taken with positive sign), in the singular vector decomposition $P = UDV^T$. Then the map $s(x_i) = v_i \quad i = 1, \dots, n$ is both strictly order-preserving and equivalence preserving.

Proof. MRP matrix P has non-negative entries, by construction, so also $P^T P$ and PP^T have; thus the Perron-Frobenius Theory applies to the Singular Value Decomposition of P (for sake of simplicity, we consider the irreducible case only, which is the more common; the non-irreducible one requires simple modifications). Without any generality loss, let $x_2 \leq x_1$ in π , so that $0 = P_{12} < P_{11} = 1$, $P_{21} = 1 = P_{22}$ and $P_{i2} \leq P_{i1}$, for all $i > 2$. From $P = UDV^T$ and the orthogonality of U , we have $V^T = D^{-1}U^T P$ and thus $v_1 = \sum_{j=1}^n u_j P_{j1}/D_{11}$ and $u_2 = \sum_{j=1}^n u_j P_{j2}/D_{11}$; as a consequence (recall that \mathbf{u} and D_{ii} are strictly positive):

$$s(x_1) - s(x_2) = v_1 - v_2 = \sum_{j=1}^n u_j (P_{j1} - P_{j2})/D_{11} > 0, \quad (1)$$

i.e. $s(x_1) > s(x_2)$.

¹ We say that x and y are equivalent in π , if there are two bijective functions ϕ and ψ from π to π , such that ϕ just exchanges x and y , ψ leaves x and y invariant and $\phi = \psi^{-1}$, i.e. one is the inverse of the other.

Again, without loss of generality, let $x_1 \sim x_2$. By the symmetry due to the equivalence relation, there exists a permutation matrix M that, in addition to switching other rows, switches the first with the second row of P and is such that $MPM^T = P$. From $(MPM^T)^T MPM^T P v = MP^T P v = D_{11} M v$, we see that the first right singular vector of $MPM^T = P$ is $M v$. By virtue of the unicity of the positive eigenvector of P , it is $M v = v$. so that $v_1 = v_2$. **q.e.d.**

By the optimal properties of the Singular Value Decomposition, we know that the first right eigenvector (up to a scale factor) provides the best unidimensional euclidean approximation to the rows of the input MRP matrix; in other words, the scores of the first right eigenvector, which is a linear combination of the rows of the MRP matrix, represent an optimal synthesis of the degree of dominance of each element of the poset on the others. The use of mutual ranking probabilities and of the spectral properties of MRP, or of other analogous matrices, to score and rank poset elements is not anew. In [3] the use of so-called *average rank* is advocated as a ranking criterion. The average rank $ark(x_i)$ of element x_i in poset π is simply the arithmetic mean of the ranks of x_i , over the set of linear extensions of π ; interestingly, it is possible to show [5] that $ark(x_i)$ is linked to mutual ranking probabilities by the following simple relation: $ark(x_i) = \sum_{j=1}^n P_{ji}$. Although the average rank is strictly order preserving and also equivalence preserving, it lacks the optimality of $s(\cdot)$ in approximating the input MRPs. In [12] and [9], it is proposed to score and rank elements by means of the first eigenvector of certain “dominance” matrices (which, for us, would coincide with P). The main drawback of using the (complex) spectral decomposition of P , instead of its singular value decomposition, is that the resulting score function is not strictly order-preserving and, when the input poset is a linear order, it scores to the same value all of the poset elements different from the maximum.

4 Smart city ranking

We use the scoring algorithm to build a “smart city” ranking, based on the 2018 study “Cities in Motion”, developed by the University of Navarra [2]². We consider the profiles composed by the ranks on three smartness pillars (*Technology, Mobility, Transportation*) of the 17 cities located in France, Germany and Italy, analyzed in the report. A poset is built according to the following partial order criterion: “city i is less smart than city j , if the former is ranked below the latter on each pillar”; Figure 1 shows the smart city poset graphically. Table 1 reports the corresponding matrix of mutual ranking probabilities and, in the last column, the scores obtained by the dominance eigenvector³ (which accounts for almost 90% of the Frobenius norm of the input MRP matrix), scaled to $[0, 1]$, by an affine transformation (table rows are

² We use this dataset just to show the procedure in action; we do not express any judgment on the conceptual model and the results of the study.

³ Computations have been performed using the R packages *parsec* [1] and *netrankr* [10].

ordered according to the final ranking). The final scores identify some subgroups of smarter or less smart cities, namely: {Paris, Berlin}, {Hamburg, Munich, Rome}, {Florence, Frankfurt, Milan}, {Cologne, Lyon}, {Lille, Marseille, Stuttgart} (which share the same score, consistently with their equivalent positions in the input poset), {Nice, Turin}, {Duisburg, Naples}.

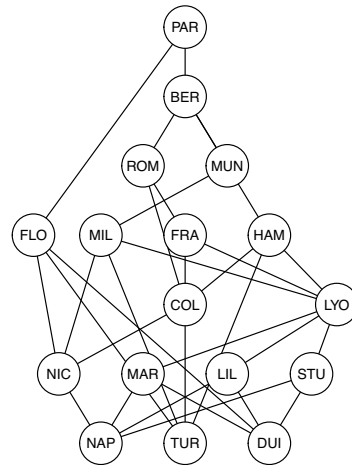


Fig. 1 Hasse diagram of the smart city poset. Two cities are connected by a downward edge, if and only if the one above is smarter, in the input poset, than the one below and there are no cities of intermediate smartness. By transitivity, one can reconstruct the entire partial order relation

5 Conclusion

In this paper we have proposed a new scoring algorithm for partially ordered data, which provides a sound procedure for ranking extraction, in a multidimensional ordinal setting, filling a gap in the data analysis literature. The procedure is computationally intensive, due to the need of getting the MRP matrix, but available software resources, and additional theoretical results [8], allow for quite efficient tools to obtain the required quantities; this makes the algorithm effective for datasets com-

	PAR	BER	ROM	MUN	HAM	FRA	MIL	FLO	LYO	COL	LIL	MAR	STU	NIC	TUR	DUI	NAP	score
PARIS	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.0000
BERLIN	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.9845
ROME	1.00	1.00	1.00	0.47	0.37	0.00	0.15	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.9079
MUNICH	1.00	1.00	0.53	1.00	0.40	0.17	0.00	0.33	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.8979
HAMBURG	1.00	1.00	0.63	0.60	1.00	0.27	0.26	0.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.8674
FRANKFURT	1.00	1.00	1.00	0.83	0.73	1.00	0.49	0.55	0.00	0.20	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.7862
MILAN	1.00	1.00	0.85	1.00	0.74	0.51	1.00	0.55	0.00	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.7842
FLORENCE	1.00	0.90	0.69	0.67	0.61	0.45	0.45	1.00	0.23	0.24	0.07	0.07	0.07	0.00	0.00	0.00	0.00	0.7787
LYON	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.77	1.00	0.50	0.00	0.00	0.00	0.09	0.12	0.00	0.00	0.6226
COLOGNE	1.00	1.00	1.00	0.94	1.00	0.80	0.79	0.76	0.50	1.00	0.20	0.20	0.20	0.00	0.17	0.02	0.00	0.5974
LILLE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	0.80	1.00	0.50	0.50	0.39	0.40	0.00	0.00	0.3639
MARSEILLE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	0.80	0.50	1.00	0.50	0.39	0.40	0.00	0.00	0.3639
STUTTGART	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	0.80	0.50	0.50	1.00	0.39	0.40	0.00	0.00	0.3639
NICE	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.91	1.00	0.61	0.61	0.61	1.00	0.49	0.15	0.00	0.2876
TURIN	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	0.83	0.60	0.60	0.60	0.51	1.00	0.22	0.18	0.2762
DUISBURG	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.85	0.78	1.00	0.42	0.0357
NAPLES	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	0.58	1.00	0.0000

Table 1 Mutual ranking probabilities matrix and final scores.

prising a much larger number of statistical units, than that used in the smart city example.

References

1. Alberto Arcagni. parsec: An R package for partial orders in socio-economics. In *M. Fattore & R. Bruggemann (Eds.) Partial Order Concepts in Applied Sciences*, pages 275–289. Springer, 2017.
2. Pascual Berrone and Joan Enric Ricart. Iese cities in motion index. *IESE Business School, University of Navarra, España*, 2018.
3. Rainer Brüggenmann and Ganapati P Patil. *Ranking and prioritization for multi-indicator systems: Introduction to partial order applications*. Springer Science & Business Media, 2011.
4. Brian A Davey and Hilary A Priestley. *Introduction to lattices and order*. Cambridge university press, 2002.
5. Karel De Loof. *Efficient computation of rank probabilities in posets*. PhD thesis, Ghent University, 2009.
6. Karel De Loof, Bernard De Baets, and Hans De Meyer. Properties of mutual rank probabilities in partially ordered sets. *Multicriteria ordering and ranking: Partial orders, ambiguities and applied issues*, pages 146–165, 2008.
7. Karel De Loof, Hans De Meyer, and Bernard De Baets. Exploiting the lattice of ideals representation of a poset. *Fundamenta Informaticae*, 71(2, 3):309–321, 2006.
8. Marco Fattore and Alberto Arcagni. A reduced posetic approach to the measurement of multidimensional ordinal deprivation. *Social Indicators Research*, 136(3):1053–1070, 2018.
9. Thomas L Saaty and G Hu. Ranking by eigenvector versus other methods in the analytic hierarchy process. *Applied Mathematics Letters*, 11(4):121–125, 1998.
10. David Schoch. *netrankr: An R package to analyze partial rankings in networks*, 2017.
11. Bernd SW Schröder. *Ordered sets: An Introduction With Connections from Combinatorics to Topology*. Springer, 2016.
12. Roberto Todeschini, Francesca Grisoni, and Serena Nembri. Weighted power–weakness ratio for multi-criteria decision making. *Chemometrics and Intelligent Laboratory Systems*, 146:329–336, 2015.

Bounded Domain Density Estimation

Stima della densità non-parametrica su domini bidimensionali limitati

Federico Ferraccioli, Laura M. Sangalli and Livio Finos

Abstract In this work we present a nonparametric penalized likelihood approach to density estimation. We consider planar domains with complex geometry, characterized by nonlinear boundaries and interior holes. The model formulation is based on a regularization with differential operators and it is made computationally tractable by means of finite element method.

Abstract *In questo lavoro viene presentato un metodo di stima di densità non-parametrica tramite verosimiglianza penalizzata. In particolare si considerano domini planari complessi, caratterizzati da bordi non lineari e buchi interni. La formulazione del modello include un termine di penalizzazione basato su operatori differenziali ed è reso computazionalmente trattabile grazie al metodo degli elementi finiti.*

Key words: Density Estimation, Penalized Likelihood, Finite Elements, Differential Operator, Permutation Test.

1 Introduction

The problem of density estimation is crucial in the statistical theory. It plays a central role in exploratory data analysis, for the visualization of structure and patterns, and

Federico Ferraccioli

Dipartimento di Scienze Statistiche, Via Cesare Battisti, 241, 35121 Padova (Italy)

e-mail: ferraccioli@stat.unipd.it

Laura M. Sangalli

MOX-Dipartimento di Matematica, Piazza L. da Vinci, 32, 20133 Milano (Italy)

e-mail: laura.sangalli@polimi.it

Livio Finos

Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Via Venezia, 8, 35131 Padova (Italy) e-mail: livio.finos@unipd.it

it may be used as intermediate procedure in classification and clustering problems. The vast majority of the research has been focused on kernel density estimation [9]. Simplicity of use and elegant analytic results are the key features of the success of the kernel approach. Bandwidth selection remains nonetheless a crucial problem for this method. Moreover, despite recent progress on the asymptotic convergence of the errors, good finite sample performance is by no means guaranteed and many practical challenges remain. In the multidimensional setting, where the specification of a symmetric positive definite bandwidth matrix is needed, the problem gets even worse, although it is common practice to use diagonal matrices.

Alongside the class of kernel density estimators, many other smoothing methods for density estimation have been proposed. The key idea of all these estimators is to reduce the complexity of the problem with some type of approximation or some form of constraint on the space of solutions. Approximations given by basis expansion with wavelets and splines are proposed in [2] and [8], respectively. Similar approaches involve regularization of the likelihood functional, such as in [7], or shape constraints on the density, e.g. log-concavity [1]. The latter is a parameter-free method but at the cost of a severe restriction on model flexibility. Regularized likelihood methods are extremely flexible; on the other hand, because of the computational complexity, they never reached popularity.

In this work we present a nonparametric penalized likelihood approach to density estimation. The model formulation is based on a regularization with differential operators and it is made computationally tractable by means of finite element method.

2 Methodology

2.1 Univariate domain: the classical approach

We introduce the problem of nonparametric maximum likelihood estimation in the univariate case, proposed for the first time in [3]. Let X_1, \dots, X_n be i.i.d. observations with distribution function F and density f on a bounded domain $\Omega \in \mathbb{R}$. Without further assumptions, the maximum likelihood estimator for f is not well defined. The estimator belongs to an infinite class of functions, resulting in an unbounded likelihood. Some type of regularization becomes necessary to avoid the trivial solution of sum of delta functions at the observations. The basic approach is to maximize a score ω , depending on f and on the observations, defined by

$$\omega = \omega(f) = L - \alpha R(f), \quad (1)$$

where $L = \sum_i \log f(x_i)$ is the log-likelihood, $R(f)$ is the roughness penalty, and the parameter $\alpha > 0$ controls the amount of smoothness. To measure the roughness or complexity of the estimate, the authors consider the functional $R(f) = \|(\sqrt{f})^{(1)}\|_2^2$. Further developments of this model are presented in [7], where a regularization functional of the form $R(f) = \|(\log f)^{(3)}\|_2^2$ is considered. In this case the limiting

estimate, as α tends to infinity, is a normal density with the same mean and variance as the data. Although both models could be in principle extended to the multivariate setting, the generalization to domain with complex shapes is far from trivial.

2.2 Bivariate domain: model and estimation procedure

We propose now a generalization of the previous model for complex planar domains. Let X_1, \dots, X_n be i.i.d. observations drawn from a distribution F with density f on a bounded planar domain $\Omega \in \mathbb{R}^2$. We define the logistic density transform $f = e^g / \int e^g$ as in [4], where g is a real function on Ω . We wish to estimate f without assuming a parametric model, by maximum likelihood estimation. As already discussed, we are dealing with some infinite class of functions. Unlike classical parametric estimation, where the parameter space is finite dimensional and the form of f is assumed known, the likelihood function is unbounded above and the degenerate solution is known to be the sum of delta functions at the observations. Some types of regularization are necessary in order to restrict the class of possible solutions. We shall consider the penalization functional given by

$$R(g) = \int_{\Omega} (\Delta g)^2 dx \quad \text{where} \quad \Delta g = \frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2}.$$

This functional measures the local curvature of g and therefore controls the smoothness of the estimates. A key feature of the Laplacian is the invariance with respect to Euclidean transformations of the spatial coordinates. It therefore ensures that the concept of smoothness does not depend on the orientation of the coordinate system. The density corresponding to the null family of the operator, i.e. when $\alpha \rightarrow +\infty$, is the uniform distribution over Ω . In the context of spatial regression model, a similar regularization approach is considered in [5].

The infinite dimensional problem can be discretized via the finite element method. Let $\psi := (\psi_1, \dots, \psi_K)^\top$ be the vector of finite element basis. We can define the discretized version of the function g as $g_D(x) = \mathbf{c}^\top \psi(x)$, where $\mathbf{c} = (c_1, \dots, c_K)^\top$ is the vector of coefficients of the finite element basis. The penalization functional is expressed by the quadratic form $\mathbf{c}^\top \mathbf{R}_1 \mathbf{R}_0^{-1} \mathbf{R}_1 \mathbf{c}$, with

$$\mathbf{R}_0 = \int_{\Omega} (\psi \psi^\top) \quad \text{and} \quad \mathbf{R}_1 = \int_{\Omega} (\psi_x \psi_x^\top + \psi_y \psi_y^\top),$$

where $\psi_x = (\partial \psi_1 / \partial x, \dots, \partial \psi_K / \partial x)^\top$ and $\psi_y = (\partial \psi_1 / \partial y, \dots, \partial \psi_K / \partial y)^\top$. For details on the derivation of the discretized regularization functional, see for instance [6].

3 Simulation study

We present a simulation study where the target density is a mixture of Gaussian distributions on the domain $\Omega = (0, 10) \times (0, 10)$. The mixture components have means $\{(2.5, 2.5), (2.5, 7.5), (7.5, 2.5), (7.5, 7.5)\}$, covariance matrices $\Sigma = \mathbb{I}_2$ and mixture weights $\pi_i = 1/4$. The density has been normalized to ensure the unity constraint on Ω . The top-left panel of figure 1 shows the contours of the function. The root mean squared error of our estimator is comparable to the KDE estimator and it has a lower variability, as shown in the top right panel of figure 1. It is important to note that our estimator has only one parameter, while the KDE needs the full covariance matrix. The two bottom panels of figure 1 show the contour of the estimates. It is interesting to note that our estimator identifies almost exactly the four mode of the density. In the case of the kernel, the estimated modes often don't match the true ones, due to the known problem of bandwidth selection.

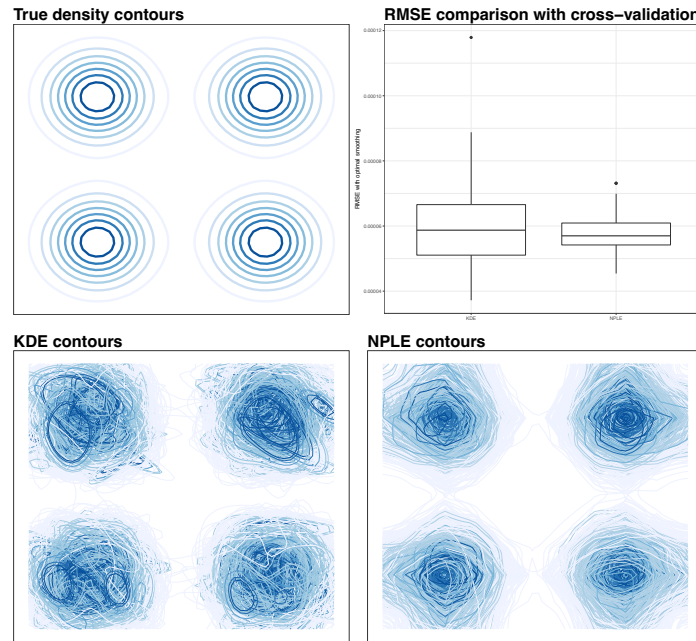


Fig. 1 Top left: contour plot of the true density. Top right: root mean squared error comparison of the two estimators. Bottom: contours of the estimated densities for the two estimators for the 100 samples.

References

1. Cule, Madeleine, Richard Samworth, and Michael Stewart. "Maximum likelihood estimation of a multi-dimensional log-concave density." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.5 (2010): 545-607.
2. Donoho, David L., et al. "Density estimation by wavelet thresholding." *The Annals of Statistics* (1996): 508-539.
3. Good, I. J., and R. A. Gaskins. "Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data." *Journal of the American Statistical Association* 75.369 (1980): 42-56.
4. Leonard, Tom. "Density estimation, stochastic processes and prior information." *Journal of the Royal Statistical Society: Series B (Methodological)* 40.2 (1978): 113-132.
5. Ramsay, Tim. "Spline smoothing over difficult regions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.2 (2002): 307-319.
6. Sangalli, Laura M., James O. Ramsay, and Timothy O. Ramsay. "Spatial spline regression models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.4 (2013): 681-703.
7. Silverman, Bernard W. "On the estimation of a probability density function by the maximum penalized likelihood method." *The Annals of Statistics* (1982): 795-810.
8. Wahba, Grace. *Spline models for observational data*. Vol. 59. Siam, 1990.
9. Wand, Matt P., and M. Chris Jones. *Kernel smoothing*. Chapman and Hall/CRC, 1994.

Polarization and long-run mobility: yearly wages comparison in three southern European countries

Polarizzazione e mobilità sul lungo periodo: un confronto fra salari annuali in tre Paesi sud-Europei

Ferretti C., Crosato L., Cipollini F., Ganugi P.

Abstract We explore the possible presence of polarization in the wage distribution for Italy, Spain and Portugal, using both descriptive statistics and the Mover Stayer model to evaluate its long-run behavior. We find that Italy and Spain are characterized by an increasing polarization, mainly due to a rise in the percentage of part-time workers, whereas Portugal maintains a unimodal wage's distribution.

Abstract *In questo lavoro si analizza il grado di polarizzazione dei salari annuali in Italia, Spagna e Portogallo, avvalendosi allo scopo sia di una analisi descrittiva sia dell'applicazione del modello Mover-Stayer. L'Italia e in particolare la Spagna sono caratterizzate da una polarizzazione che cresce nel lungo periodo, principalmente per effetto dell'aumento percentuale dei lavoratori part-time. Lo stesso fenomeno non riguarda il Portogallo, che mantiene una distribuzione unimodale.*

Key words: Wage distribution, polarization, inequality, mobility

1 Introduction

Wage distribution has been extensively analysed in terms of its relationship with the job structure, as in [14], or looking for a suitable quantification of the degree of polarization as in [8, 9]. Polarization, intended as the presence of two or more

Ferretti Camilla
Univ. Cattolica del Sacro Cuore, Piacenza, e-mail: camilla.ferretti@unicatt.it

Crosato Lisa
Univ. degli Studi di Milano-Bicocca e-mail: lisa.crosato@unimib.it

Cipollini Fabrizio
Univ. degli Studi di Firenze e-mail: cipollini@disia.unifi.it

Ganugi Piero
Univ. degli Studi di Parma e-mail: piero.ganugi@unipr.it

extreme poles in the distribution around which individuals tend to cluster, has been observed during the last decades in the U.S. labor market and explained by cross-country technological catch-up ([7]) or by routine-biased technical change ([6]). Little evidence of wage polarization in Germany was instead found by [2].

In this paper, we contribute to the existing literature by investigating the distribution of the employees' yearly net income in Italy, Portugal and Spain using EUsilc data¹. We compare the trends in polarization and inequality in the three countries, finding a similar evolution for Italy and Spain, and an opposite behaviour in Portugal.

In addition, and differently from previous contributions, we focus also on the long-run evolution of individuals among the wage categories estimating the parameters of a stochastic process called Mover Stayer (MS), which permits to forecast the tendency of workers to remain in their initial category as well as the long-run shape of the distribution. It is a first step in addressing the issue of a possible relationship between mobility and polarization, being aware that the halt of upward mobility across earning ladders in the last decade and the simultaneous increasing of inequality and polarization has been object of several studies, as for example [15]: among others, [3] finds that wage inequality is reduced by mobility, although with large differences across European countries; on the other hand, [16] observes that in Germany inequality and wage mobility raised at the same time.

2 Data description and descriptive statistics

Data are extracted from the EUsilc database: we consider waves 2008 and 2016 focusing on the variable PY010N (*Employee cash or near cash income*). All the values are rescaled to take into account the differences in the price levels² (2015 = 1) and in the PPP³ (EU28 = 1). For the category-based analyses, we consider ten wage groups based on the 2005-deciles, calculated aggregating data of all European Countries having at least 1000 valid observations in that year (cf. Table 1). Zero values are excluded.

The PY010N observations are stored in the longitudinal component of EUsilc and derive from yearly interviews of a rotating sample of individuals. Since a quarter of the sample is renewed each year, there exists a sub-sample of individuals which are included for four consecutive years. Accordingly, we considered two four-years-balanced panels corresponding to the periods 2005 - 2008 (from wave 2008) and 2013 - 2016 (from wave 2016). All the following analyses are based on such panels.

¹ EUsilc, 2005-2016, data are obtained with the research proposal no 92/2017. The responsibility for all conclusions drawn from the data lies entirely with the authors.

² Harmonized consumer price index HICP: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=prc_hicp_aind&lang=en

³ Purchasing power parity index: <https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tec00120&plugin=1>

Table 1 Equivalized yearly wage categories based on the European 2005-deciles (values are expressed in Euros at 2015 value with purchasing power parity EU28 = 1).

I	II	III	IV	V
(0, 3900]	(3900, 7650]	(7650, 11600]	(11600, 14450]	(14450, 16800]
VI	VII	VIII	IX	X
(16800, 19000]	(19000, 22250]	(22250, 26300]	(26300, 33350]	(33350, +∞)

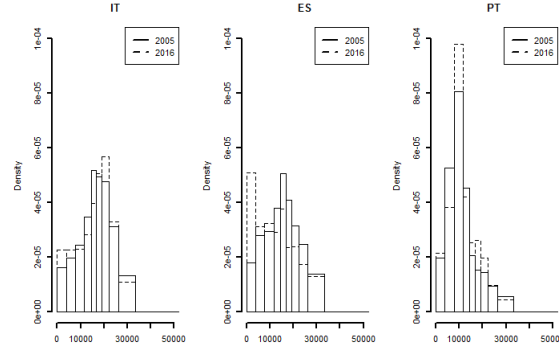
Fig. 1 Wage distributions for Italy, Spain and Portugal in 2005 (solid line) and in 2016 (dashed line). Categories are defined in Table 1. Panels for Italy, Spain and Portugal contain resp. 8451, 5382 and 1880 individuals in the years 2005-2008 and 6804, 4909 and 3253 individuals in the years 2013-2016.

Figure 1 shows that in 2005 all the distributions are unimodal. In 2016 Portugal maintains this characteristic whereas Spain shows a first mode in the leftmost category and another lower mode in the intermediate part; Italy lies in an intermediate situation, with a substantially flat shape in the first three categories.

It is interesting to link this evolution of wages to the degree of inequality and polarization. Inequality is evaluated by the classical Gini index, polarization is calculated as in [9], using the DER index:

$$DER = \frac{1}{n} \sum_{i=1}^n \hat{f}(w_i) \hat{\alpha}(w_i),$$

where w_1, \dots, w_n are ordered observed wages, $\hat{f}(w_i)$ is a density estimate in w_i obtained through the kernel method and $\hat{\alpha}(w_i) = \bar{w} - w_i(n^{-1}(2i-1) - 1) - n^{-1}(\sum_{j=1}^{i-1} w_j + w_i)$. Polarization indices check for the possible presence of two or more distinct groups in the data. The higher is the index, the higher is the possibility to split the distribution in many distinct clusters.

As shown in Table 2 inequality and polarization have a similar behaviour: they decrease in Portugal and increase in Italy and Spain. In addition, the DER index behaves as expected, for example increasing in Spain in the last years. Nevertheless, neither Gini nor DER index neatly discriminate between Portugal and the other countries, which show very similar values.

Table 2 Gini index of concentration in comparison with DER index of polarization.

	GINI			DER		
Year	IT	ES	PT	IT	ES	PT
2005	0.296	0.337	0.411	0.386	0.418	0.493
2008	0.316	0.341	0.375	0.405	0.420	0.459
2013	0.352	0.420	0.324	0.436	0.483	0.416
2016	0.310	0.444	0.318	0.403	0.503	0.413

3 Long-run polarization

The descriptive analysis presented in Section 2 provides a snapshot of the wage distributions in some selected years. To enrich such analysis we propose to make some inference about the predicted time evolution of wages. To this aim we assume that movements of individuals among the categories defined in Table 1 are ruled by a discrete-time Mover Stayer model (see [11] and [13], among others) which is a mixture of two Markov chains and is based on the hypothesis that there exists a group of individuals never moving from their starting category. Consequently, yearly transitions of individuals among the categories are supposed to be given by the 10×10 transition matrix $P = S + (I - S)M$, where S is a diagonal matrix containing the probabilities to be a Stayer for every starting category, and M is the classically-defined transition matrix for the non Stayers individuals.

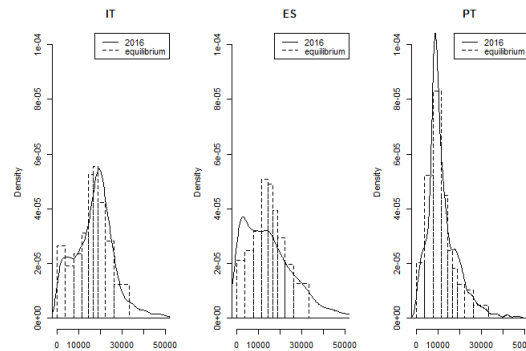
The MS model offers additional elements for evaluating the wage mobility: on one hand it provides the probability to be a Stayer in every category, namely a worker able to resist economic changes; on the other hand it allows us to estimate the long-run distribution, that is the distribution to which workers would tend if the observed economic situation remained stable.

S and M matrices are estimated on the basis of the balanced panel 2005 - 2008 with the aim to evaluate how wage distribution would be if the 2008 crisis never happened.⁴ For sake of space, Table 3 reports only the estimated S together with the corresponding standard error. The comparison between the estimated equilibrium based on the 2005-2008 data and the 2016 observed distribution is shown in Figure 2. It is relevant to note that the probability to be a Stayer is high (around 40%), as expected, in the last wage category (yearly wage > 33350 equivalized Euros). Nevertheless Spanish workers generally show a lower probability to resist the economic changes than Italy and Portugal, especially for categories VIII and IX (yearly wage comprised between 22250 and 33350 equivalized Euros). Spain also has a peculiar behaviour in terms of the expected long-run distribution: we note that the 2016 observed distribution is more polarized than expected according with pre-crisis data. On the other hand, wage distributions in Italy and Portugal experienced a weaker change in their shape.

⁴ Estimation is based on a Bayesian approach according to [5] and using OpenBUGS (5000 burn-in iterations and 5000 MCMC iterations with a lag equal to 50 to avoid autocorrelation).

Table 3 Estimated probabilities (expressed as percentages) to be a Stayer by wage category, and the corresponding standard error (in italic) obtained using the 2005 - 2008 balanced panels.

Class	I	II	III	IV	V	VI	VII	VIII	IX	X
IT	2.66	6.21	9.45	7.04	11.90	10.94	13.82	16.86	23.70	41.98
<i>se</i>	<i>1.25</i>	<i>1.75</i>	<i>2.02</i>	<i>1.67</i>	<i>2.28</i>	<i>2.05</i>	<i>1.99</i>	<i>2.32</i>	<i>2.92</i>	<i>4.25</i>
ES	1.81	3.44	7.68	5.83	3.49	1.86	8.87	5.29	4.25	37.80
<i>se</i>	<i>1.31</i>	<i>1.49</i>	<i>1.99</i>	<i>2.11</i>	<i>1.55</i>	<i>1.21</i>	<i>2.05</i>	<i>1.88</i>	<i>2.33</i>	<i>4.18</i>
PT	5.99	12.24	22.91	5.41	5.17	7.32	5.37	11.69	10.42	43.56
<i>se</i>	<i>3.59</i>	<i>4.88</i>	<i>5.10</i>	<i>3.67</i>	<i>3.96</i>	<i>5.23</i>	<i>4.29</i>	<i>7.72</i>	<i>6.91</i>	<i>4.26</i>

Fig. 2 Observed 2016 distribution (kernel density, solid line) and the long-run distribution (dashed histograms) expected from wave 2005-2008.

4 Discussion

In the previous sections we analyzed the yearly wage distributions without taking into account the time really worked by each person in the year. On the same line, the observed polarization is actually the result of a mixture of separated distributions related to the type of employment (full-time, part-time and other) as Table 4 shows.

Table 4 Percentages of workers having a full-time/part-time/other employment in 2005 and 2006.

	Italy			Spain			Portugal		
Year	Full-time	Part-time	Other	Full-time	Part-time	Other	Full-time	Part-time	Other
2005	73%	10%	17%	80%	8%	12%	87%	5%	8%
2016	69%	14%	17%	62%	12%	26%	86%	5%	9%

From the same table we note that the percentage of full-time contracts in Portugal remains stable in 2016 with respect to 2005, whereas it decreases in Italy and Spain. This behaviour is coherent with the results previously displayed and evidence just discussed may highlight an increasing presence of non voluntary part time contracts imposed by firms owners to contrast the shortage in production after the 2008 Great recession. Accordingly, for future research we propose to investigate how "weak"

workers (part-time with a low wage) evolve in the time and if they move towards a full-time contract, self-employment or unemployment.

With the same purpose (the analysis of the capacity of "weak" and "strong" workers to resist economic changes), we will investigate if it is possible to characterize the probability to be a Stayer in terms of personal covariates, using a refined version of the logistic regression as in [12].

In terms of the polarization analysis, we will follow the line of [1] which points out the presence of two poles in the distribution of gross hourly wages in Spain and Italy and attributes it to the characteristics of the individuals in the labor market. Lastly, we aim to examine the relationship among polarization and mobility indices, measured for example as in [10].

References

1. Addabbo, T., García-Fernández, R.M., Llorca-Rodríguez, C.M. Maccagnan, A.: Labor force heterogeneity and wage polarization: Italy and Spain. *J. Econ Stud* **45:5**, 979–993 (2018)
2. Antonczyk, D., DeLeire, T., Fitzenberger, B.: Polarization and Rising Wage Inequality: Comparing the US and Germany. *Econometrics* **6:2**, 1 - 33 (2018)
3. Bachmann, R., Bechara, P., Schaffner, S.: Wage Inequality and Wage Mobility in Europe, *Rev Income Wealth* **62:1**, 181–197 (2016)
4. Castellano, R., Musella, G., Punzo, G.: Structure of the labour market and wage inequality: evidence from European countries. *Quality & Quantity* **51:5**, 2191–2218 (2017)
5. Cipollini, F., Ferretti, C., Ganugi, P., and Mezzanzanica M.: A Continuous Time Mover-Stayer Model for Labor Market in a Northern Italian Area. *Classification and Data Mining* 1018 - 188 (2013) DOI 10.1007/978-3-642-28894-422
6. Cortes, G.M.: Where have the middle-wage workers gone? a study of polarization using panel data. *J Labor Econ* **34:1**, 63–105 (2016)
7. Cozzi, G., Impulliti, G.: Globalization and wage polarization. *Rev Econ Stat* **98:5**, 984–1000 (2016)
8. Deutsch, J., Fusco, A., Silber, J.: The BIP trilogy (Bipolarization, Inequality and Polarization): one saga but three different stories. *Econ* **7**, 1 -33 (2013)
9. Duclos, J.Y, Esteban, J., Ray, D.: Polarization: Concepts, Measurement, Estimation. *Econometrica* **72:6**, 1737 - 1772 (2004)
10. Ferretti, C., Ganugi, P.: A new mobility index for transition matrices. *Stat Methods Appl* **22**, 403 - 425 (2013)
11. Fougere, D., Kamionka, T.: Bayesian inference for the Mover-Stayer model of continuous time. *J App Econ* **18** 697723 (2013)
12. Frydman, H., Matuszyk, A.: Estimation and status prediction in a discrete mover-stayer model with covariate effects on stayer's probability. *App Stoch Mod Bus Ind* **33:5** 1 - 10 (2017)
13. Goodman, L.A.: Statistical methods for the Mover Stayer model. *J Am Stat Ass* **56** 841 - 868 (1961)
14. Goos, M., Manning, A.: Lousy and Lovely Jobs: the rising polarization of work in Britain. *Rev Econ Stat* **89:1**, 118 - 133 (2007)
15. Poggi, A., Silber, J.: On polarization and mobility: a look at polarization in the wage-career profile in Italy. *Rev Income Wealth* **56:1**, 123 - 140 (2010)
16. Riphahn, R.T., Schnitzlein, D.D.: Wage mobility in east and west Germany. *Labour Econ* **39**, 11–34 (2016)

Design of Experiments, aberration and Market Basket Analysis

Pianificazione degli esperimenti, aberrazione e Market Basket Analysis

Roberto Fontana and Fabio Rapallo

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

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

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

1 Introduction

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

Roberto Fontana

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

Rapallo Fabio

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

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

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

2 Fractional factorial designs and aberrations

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and δ_a^b denotes the Kronecker delta.

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

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

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

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

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

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

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

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

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

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

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

3 An application to Italian museums

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

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

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

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

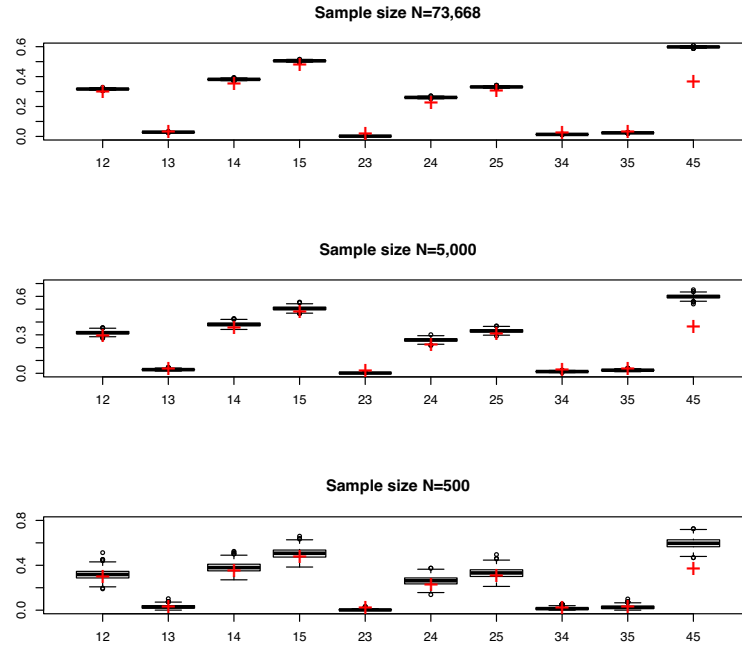


Fig. 1 Distribution of the aberrations of order two for different sample sizes. The symbol ‘+’ denotes the expected aberration under independence.

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

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

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

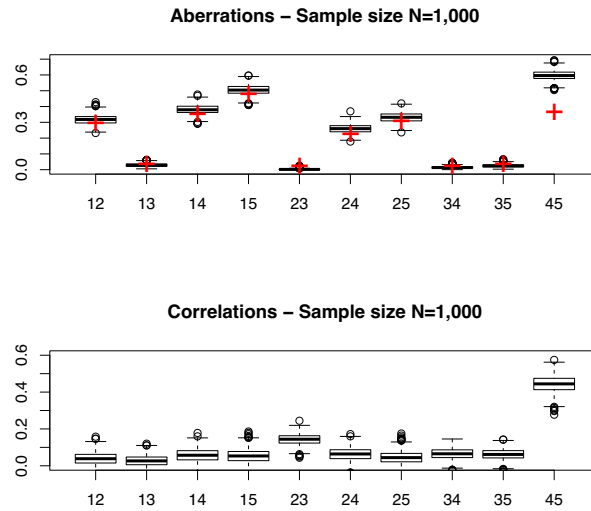


Fig. 2 Distribution of the aberrations of order two and the corresponding correlation coefficients.

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

References

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

Generalized Procrustes Analysis for Multilingual Studies

Analisi Procustiana Generalizzata per studi Multilingue

Alessia Forciniti, Michelangelo Misuraca, Germana Scepi, Maria Spano

Abstract When the aim is to study the differences among cultures, the comparative cultural studies assume considerable importance. One of the most peculiar aspects of a culture is definitely the language as a medium of communication between people. In the literature of cross language analyses, many approaches have been proposed for comparing two languages. They are mainly based on use of bilingual dictionaries and machine translation systems. In order to compare simultaneously more languages, we propose to use the Generalized Procrustes Analysis on Twitter data for analysing the cross-country public opinion. A case study about the violence against women is presented for showing the effectiveness of our proposal.

Abstract *Quando lo scopo è studiare le differenze tra culture, gli studi culturali comparativi assumono considerevole importanza. Uno degli aspetti più peculiari di una cultura è sicuramente la lingua come mezzo di comunicazione tra persone. Nella letteratura di analisi interlinguistiche, sono stati proposti molti approcci per il confronto di due lingue. Si basano principalmente sull'uso di dizionari bilingue e sistemi di traduzione automatica. Al fine di confrontare simultaneamente più lingue, proponiamo l'Analisi Procustiana Generalizzata su dati Twitter per analizzare l'opinione pubblica tra Paesi. Viene proposto un caso-studio sulla violenza contro le donne per mostrare l'efficacia della nostra proposta.*

Key words: Generalized Procrustes Analysis, GPA, Cross-language studies, Textual data, Violence against Women

¹ Forciniti A, University of Naples Federico II; email: alessia.forciniti@unina.it

² Misuraca M, University of Calabria; email: michelangelo.misuraca@unical.it

³ Scepi G, University of Naples Federico II; email: germana.scepi@unina.it

⁴ Spano M, University of Naples Federico II; email: maria.spano@unina.it

1 Introduction

In a globalized social context, the comparative cultural studies (Tötösy de Zepetnek, S., Tutun, M., (2013)) assume considerable importance with the objective to study the differences between cultures and culture products in a transnational perspective. There are several elements characterizing a culture and the most peculiar aspect is the language. Generally, the cross-country analyses are performed with a plurality of methods in according to the research objectives. If the focus is to study the natural language expressions among different countries with the purpose of highlighting their similarities or dissimilarities, we talk about *cross-languages* methods. Various methodological approaches have been developed in literature. However, many of these compare just two languages and/or *parallel corpora*, where a *corpus* is the exact translation to another one. Despite different contributions for the cross-lingual studies, our interest moves towards an analysis that involves more languages at the same time and focuses on non-parallel corpora. About this, we propose the *Generalized Procrustes Analysis*, which permits to represent more factorial configurations in a common geometrical space; typical in the sensorial sciences for comparing the differences in scoring between the individuals (say judges) about products and or variables of these (Williams and Langron, (1984); Arnold and Williams, (1986)). In this case, it is adopted for comparing simultaneously the differences in public opinion on Twitter data, in four different languages. In the following, in section 2 is briefly introduced the theoretical framework of cross language analyses and then the methodological approach used in our strategy. In section 3, we present a case study on the phenomenon of violence against women.

2 Theoretical framework

In literature, there is a considerable proposal of cross-language methods. Mihalcea *et al.* (2007) study the multilingual subjectivity between two lingual cultures by using bilingual dictionaries. One of the languages for the comparison is English. Furthermore, in most cases, the analyses are conducted on parallel corpora. The study of Vinokourov *et al.* (2003) indicates the use of the Kernel Canonical Correlation Analysis for the comparison between two parallel corpora: English and French. Similarly, the Mihalcea *et al.*'s investigation (2007) proposes a cross-language strategy on parallel corpora for developing a statistical classifier. For a cross-lingual alignment of parallel corpora between Hebrew and Arabic, Snyder and Barzilay (2008) present non-parametric Bayesian models by separating the lexical unit into its individual morphemes. Several approaches explore the use of a third language: these studies are performed with English as training and testing language. Nagaraja and Jayanna (2012) analyse through monolingual and cross- lingual

methods the speaker identification system in communication-transactions field, by considering English, Hindi and Kannada languages. The research is performed with English as training and testing language. Other approaches are based on the support of machine translation systems. The work of Banea *et al.* (2008) evaluates the transferring of subjectivity through the use of English as third language and considers the machine translation systems of other languages with respect to English. If we do not consider the methods up to here mentioned, for comparing simultaneously more languages and non-parallel corpora, we propose the *Procrustes Analysis* on $m=4$ configurations of multilingual co-occurrence matrices.

2.1 Generalized Procrustes Analysis

In 1975, Gower introduces a multivariate exploratory technique called *Generalized Procrustes Analysis* (GPA) as development of *Classical Procrustes Analysis* (PA; Hurley, Cattell, (1962)) for comparing simultaneously several sets of data or configurations and returning a group average configuration (Dijksterhuis, Gower, (2004)). The configurations identified as m (where $m \geq 2$), usually, are two-dimensional matrices $n \times p$, where n are individuals or objects, into a p -dimensional space of variables. Suppose $P_j^{(i)} (i=1,2,\dots,m, j=1,2,\dots,n)$ as mn points in a p -dimensional space; where we have m ($n \times p$) matrices $X_i (i=1,2,\dots,m)$ and where the j^{th} row of X_i is interpretable as coordinates of a point $P_j^{(i)}$ referred to p_i orthogonal axes. The GPA with the aid of appropriate transformations: translation, scaling, rotation-reflection provides a goodness of fit criterion for an optimal comparability. From a geometrical point of view, the criterion permits to represent, m sets of matrices in a common space by minimizing the sum-of-squares between mn points and their centroid (G_j). The distances between $P_j^{(i)}$ and G_j are denominated *residuals*. The average configuration represents the *consensus* while the relationship with individual scaling, in computational terms, can be summarized through the variance (SS). For improving the interpretability of the consensus among the mn points and for estimating their proportion of variability, Gower suggests, in his work of 1975, to combine the GPA graphic visualization with the analysis of variance. The analysis of variance is generalized (ANOVA) and for GPA is denominated Procrustes Generalized Analysis of Variance (PANOVA; for its interpretation refer to Commandeur, (1991)). Wakeling *et al.* (1992) develop a permutation test for checking if the consensus obtained by GPA is meaningful. This test determinates if the observed *Rc value* (corresponding to the proportion of the variance explained by the consensus configuration) is meaningful by considering $p\text{-value} < .05$.

3 A case study

GPA is applied to Twitter multimedia content for mapping the public opinion in relation to violence against women in four European Countries. The countries:

Czech Republic, France, Italy and United Kingdom are chosen in relation to the Welfare State Theory of Esping-Andersen (1999). Data extraction is performed through web scraping techniques from 21st to 29th November 2018; during the week around the International Day for the Elimination of Violence against Women (25th November). By using three keywords: “*violence*”, “*women*” and “*international day*” in four different languages and by geocoding the Countries, we collected 5432 tweets. The pre-processing step consists in normalizing the text, removing URL, special characters, emoticons and stop words for each language. In data analysis phase, we consider only common terms among countries that occur with a threshold greater/equal to 10. In addition, the words with similar meaningful are grouped into just one semantic category equal for each linguistic configuration. Therefore, we have a common vocabulary for all countries: four *document features matrices* and the corresponding *features x features co-occurrence matrices* of 16×16 for each language. These are adjacency tables in which, each element represents the number of times two terms co-occur in the collection of tweets and it can be used for analysing the relations existing among the different *features*.

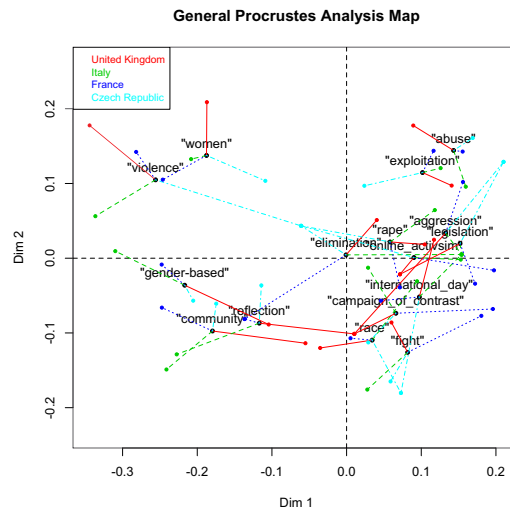


Figure 1: Geometry of match for data

At the end of pre-processing step, we consider $m=4$ (configurations and so the countries), $p=16$ (common occurrences) and $n=16$ (each occurrence is located in relation to the correspondent translated term). For saying as Gower (1975) the objects of analysis are the common features (for interpretative convenience are presented in English), the countries represent the judges and the attributes of objects are the translated occurrences. The GPA Map, shows the position, in the geometrical space of the common terms used by Twitter users with respect to the centroid of each lexical unit (G_1, \dots, G_{16}). Each segment represents a country. The *Figure 1*

shows an interesting visualization about the term “*violence*”: France is the closest configuration respect to average point while Czech Republic is the most distant. The most reduced distance among all linguistic configurations is in first quadrant, for the term “*exploitation*”, since the trajectories of the countries are very close to the centroid by denoting similarities among them. For demonstrating these graphic considerations and helping the interpretation, the map is studied together to PANOVA as suggested by Gower (1975).

Table 1: Analysis of Variance (PANOA)

<i>Terms</i>	<i>SS_{fit}</i>	<i>SS_{residual}</i>	<i>SS_{total}</i>
violence	10.0318298	2.5909285	12.622758
women	7.6140839	0.9716165	8.585700
gender-based	7.5834278	1.2034198	8.786848
community	5.9340330	1.1488966	7.082930
reflection	4.1454543	1.6668599	5.812314
elimination	0.7442335	2.5028352	3.247069
race	4.7007783	1.4299118	6.130690
international_day	3.6235972	2.0071846	5.630782
fight	5.5022715	1.2562297	6.758501
campaign_of_contrast	3.3306129	1.9416216	5.272235
online_activism	1.8959138	1.7764544	3.672368
legislation	3.7492894	1.7918447	5.541134
aggression	3.0822329	2.5654614	5.647694
abuse	6.1069271	1.0536390	7.160566
exploitation	4.2657788	0.4688275	4.734606
rape	2.0735259	1.2402789	3.313805
Sum	74.3839901	25.6160099	100.000000

<i>Configurations</i>	<i>SS_{residual}</i>	<i>SS_{total}</i>
United Kingdom	7.007152	23.76323
Italy	5.356221	27.14850
France	5.417206	27.02345
Czech Republic	7.835432	22.06482
Sum	25.616010	100.000000

PANOVA (Table 1) indicates that the total fit and hence the consensus of this multilingual study is equal to 0.7438, that is, the 74% of the sum of squares of the individual configurations. Therefore, GPA model fits well to the data; by suggesting that there is a common structure in the way to percept the phenomenon between the four investigated countries and there is a similarity in the grammatical structure of the multimedia contents. As the graphic visualization, “*exploitation*” indicates the best fitting with the lowest residuals; taking into account that more are high the residuals and more differ the objects, in terms of consensus. To confirm the graphic interpretation about the linguistic trajectories, the lowest matching comes from

“*violence*” presenting the highest residuals. The words more fitted are: “*exploitation*”, “*women*” and “*abuse*”; while the less performative are “*violence*”, “*aggression*” and “*elimination*”. The variance of the $m=4$ configurations, indicates the highest residuals in respect to Czech Republic and the lowest residuals in relation to French and the Italy. Therefore, French and Italy show higher similarity. GPA-test indicates the proportion of the original variance explained by the consensus configuration by means of the R_c value. This is of 74% with significance of 0.001. GPA on the multilingual co-occurrence matrices represents a method useful for cross-language studies, since permits to explore similarities or dissimilarities among linguistic configurations and among several cultures. In other words, the distances among lexical units from their average configurations (consensus) provide two different perspective of analysis. The first compares the syntactic structure among the languages through the study of relations existing among the different features. This evaluates the linguistic similarity in terms of the sentences structure. The second perspective identifies the language as bearer of values and beliefs and therefore investigates a common structure of sense in the way to percept phenomena, by showing cultural similarities among countries.

References

1. Arnold, G.M., Williams, A.A.:The Use of Generalized Procrustes Techniques in Sensory Analysis. In: Piggot, J.R., Ed., Statistical Procedures in Food Research, Elsevier, London, 233-253 (1986)
2. Banea,C., *et al.*: Multilingual subjectivity analysis using machine translation. In EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2008)
3. Council of Europe:Council of Europe Convention on preventing and combating violence against women and domestic violence (2011)
4. Dijksterhuis,G.B., Gower, J.C.:Procrustes Problems, Oxford Statistical Social Science Series (2004)
5. Esping-Andersen,G.:Social Foundations of Postindustrial Economies, Oxford Univ. Press (1999)
6. Gower, J.C.: Generalized Procrustes analysis, Psychometrika, 40, 33-50 (1975)
7. Hurley, J.R.,Cattel, R., B.:The procrustes program: Producing direct rotation to test a hypothesized factor structure. Computers in Behavioral Science, Journal of the Society for General System Research, Vol.7, Issue 2, pp. 258-262 (1962)
8. Mihalcea, R., Banea, C., Wiebe, J.: Learning Multilingual Subjective Language via Cross-Lingual Projections.Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 976-983, Prague, Czech Republic (2007)
9. Nagaraja, B.G., Jayanna, H.S.: Mono and Cross lingual speaker identification with the constraint of limited data. International Conference on Pattern Recognition, Informatics and Medical Engineering, PRIME-2012 (2012)
10. Snyder, B., Barzilay, R.:Cross-lingual Propagation for Morphological Analysis. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008).
11. Tötösy de Zepetnek, S., Tutun, M.: eds. Companion to Comparative Literature, World Literatures, and Comparative Cultural Studies. New Delhi: Cambridge UP, India (2013)
12. Vinokourov, A. et al.: Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. Advances in neural information processing systems (2003)
13. Wakeling, I.N., Raats, M.M., MacFie, H.J.H.:A new significance test for consensus in Generalized, Journal of Sensory Studies 7:91-96 (1992)
14. Williams, A.A., Langron, S.,P.:The Use of Free-Choice Profiling for the Evaluation of Commercial Ports. Journal of the Science of Food and Agriculture 35(5): 558-568 (1984)

Prior specification in flexible models

Specificazione delle prior in modelli flessibili

Maria Franco-Villoria, Massimo Ventrucchi and Haavard Rue

Abstract The linear predictor of generalized additive models is expressed as a sum of unspecified smooth functions. In a Bayesian hierarchical framework, smooth functions can be described by a vector of random effects distributed at prior as a Gaussian Markov Random Field. In this work, we present the use of Penalized Complexity Priors (PC priors) for flexible models, introducing a natural base model.

Abstract *Il predittore lineare nei modelli additivi generalizzati viene espresso come una somma di funzioni smooth (i.e. forma parametrica non specificata). Nel contesto dei modelli gerarchici Bayesiani, le funzioni smooth vengono descritte da un vettore di effetti casuali distribuiti a priori come un Gaussian Markov Random Field. In questo lavoro rivisitiamo i modelli flessibili attraverso l'uso della classe di distribuzioni a priori nota come Penalized Complexity Priors (PC priors).*

Key words: base model, Gaussian Markov random field, penalized complexity, random walk

1 Introduction

In some practical applications imposing a linear relationship between the response and explanatory variable might be too restrictive, and more flexible models, such as generalized additive models (GAM) [6], might be needed. In these flexible models, the linear predictor is specified as a sum of smooth functions of the explanatory vari-

Maria Franco-Villoria
University of Torino e-mail: maria.francovilloria@unito.it

Massimo Ventrucchi
University of Bologna, e-mail: massimo.ventrucchi@unibo.it

Haavard Rue
King Abdullah University of Science and Technology e-mail: haavard.rue@kaust.edu.sa

ables. We follow a Bayesian hierarchical framework where each smooth function is described by a vector of random effects distributed at prior as a Gaussian Markov Random Field (GMRF) [4]. A GMRF is a multivariate normal distribution with mean vector μ and a sparse precision $Q(\tau)$ that depends on some hyper-parameters τ and whose non zero pattern specifies conditional dependencies among neighbouring random effects.

Elicitation of priors for precision parameters is a long standing topic in the literature on hierarchical Bayesian models. Simpson et al. (2017) [5] recently introduced a new framework for building priors that avoid overfitting denoted as *Penalized Complexity (PC) priors*. PC priors are computed based on specific principles in which a model component is seen as a flexible parametrization of a base model. The idea is to penalize model complexity, defined in terms of distance from the base model, in such a way that the base model is favoured unless the available data support a more flexible one.

2 Penalized Complexity (PC) Priors

In this section we summarize the four main principles underpinning the construction of PC priors, namely: support to Occam's razor (parsimony), penalisation of model complexity, constant rate penalisation and user-defined scaling. For a more detailed presentation of these principles the reader is referred to [5].

Let f_1 denote the density of a model component w where τ is the parameter for which we need to specify a prior. The base model, corresponds to a fixed value of the parameter $\tau = \tau_0$ and is characterized by the density f_0 .

1. The prior for τ should give proper shrinkage to τ_0 and decay with increasing complexity of f_1 in support of Occam's razor, ensuring parsimony; i.e. the simplest model is favoured unless there is evidence for a more flexible one.
2. The increased complexity of f_1 with respect to f_0 is measured using the Kullback-Leibler divergence (*KLD*) [2],

$$KLD(f_1||f_0) = \int f_1(w) \log \left(\frac{f_1(w)}{f_0(w)} \right) dw,$$

which, for zero mean multivariate normal densities is

$$KLD(f_1||f_0) = \frac{1}{2} \left(tr(\Sigma_0^{-1} \Sigma_1) - n - \ln \left(\frac{|\Sigma_1|}{|\Sigma_0|} \right) \right)$$

where n is the dimension. For ease of interpretation, the KLD is transformed to a unidirectional distance measure

$$d(\tau) = d(f_1||f_0) = \sqrt{2KLD(f_1||f_0)} \quad (1)$$

that can be interpreted as the distance from the flexible model f_1 to the base model f_0 .

3. The PC prior is defined as an exponential distribution on the distance, $\pi(d(\tau)) = \lambda \exp(-\lambda d(\tau))$, with rate $\lambda > 0$, ensuring constant rate penalization. Therefore, the mode of a PC prior is always at the base model. The PC prior for τ follows by a change of variable transformation.
4. The user must select λ based on his prior knowledge on the parameter of interest (or an interpretable transformation of it $T(\tau)$). This knowledge can be expressed in terms of a probability statement, e.g. $P(T(\tau) > U) = a$, where U is an upper bound for $T(\tau)$ and a is a (generally small) probability.

One major advantage of PC priors is that they prevent overfitting by construction, as they guarantee shrinkage towards the base model. PC priors for the marginal variance of a Gaussian random effect have been shown to outperform other priors widely used in literature (such as Inverse Gamma priors) when data are weakly informative or the size of the effects is close to the base model [1]. Finally, prior information, if available, can be coded into an intuitive way by simply specifying U and a .

3 Flexible models

Consider the simple case where there are n observational units indexed by $i = 1, \dots, n$ and one covariate x_i whose effect on the response y_i is not assumed to take any parametric shape. Assuming y_i belonging to the exponential family, the linear predictor of a generalized additive model is

$$\eta_i = \alpha + f(x_i) \quad i = 1, \dots, n. \quad (2)$$

The smooth function in (2) can be described by a vector of random effects $\beta = (\beta_1, \dots, \beta_n)^T$, for which random walk priors are a very popular choice. A random walk is a particular intrinsic GMRF of order r , i.e. a process characterized by the following improper multivariate Gaussian distribution:

$$\pi(\beta|\tau) = (2\pi)^{-rank(R)/2} (|\tau R|^*)^{1/2} \exp \left\{ -\frac{\tau}{2} \beta^T R \beta \right\} \quad (3)$$

where $|\tau R|^*$ is the generalized determinant. Density (3) is improper as it is invariant to the addition of a polynomial of degree $r - 1$. In the case of a random walk of order 2, a useful parametrization for the smooth function in Model (2) is

$$f(x_i) = \beta_0 x_i + \beta_i, \quad i = 1, \dots, n \quad (4)$$

where β_i are subject to the linear constraints $\sum_{i=1}^n \beta_i = 0$, $\sum_{i=1}^n x_i \beta_i = 0$ and can be seen as deviations from the linear trend $\beta_0 x_i$. The model turns into a simple linear regression model when the smooth function $f(x_i)$ is linear over x_i , i.e. when $\beta_i = 0$

$\forall i$. The linear model can be regarded as a *base model*, while Model (2) can be seen as a flexible extension of it. The precision parameter τ controls how flexible the corresponding smooth function is, as shown in Fig. 1. The base model can be obtained setting the hyper-parameter $\tau = \infty$.

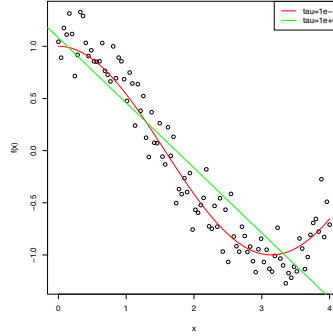


Fig. 1 Effect of the precision parameter on fitted smooth function assuming a RW2 prior.

4 Work in Progress

We consider rewriting Model (2) using an alternative and intuitive parametrization following the work by [3]. This way prior distributions for β_0 and β_i in (4) can be set jointly. The parametrization considered can be easily extended for models involving bivariate smooth functions.

References

1. Klein, N. and Kneib, T. (2016). Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression. *Bayesian Analysis*, 11(4):1071–1106.
2. Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.
3. Riebler, A., Srbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145–1165. PMID: 27566770.
4. Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman and Hall/CRC.
5. Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Srbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.*, 32(1):1–28.
6. Wood, S.N (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Modeling Cyclists' Itinerary Choices: Evidence from a Docking Station-Based Bike-Sharing System

Un modello per gli itinerari dei ciclisti: risultati da un bike-sharing a stazioni fisse

S. T. Gaito - G. Manzi - G. Saibene - S. Salini - M. Zignani

Abstract This paper presents a bike itinerary choice model for a bike sharing system (BSS). Machine learning techniques are used to characterize the bike itineraries within a given day. We present a method to detect the 'most faithful' users with respect to itinerary types and a Bayesian network model which emphasizes the link between the conditions with which the itinerary is chosen and the itinerary type.

Abstract *Questo articolo presenta un modello di scelta di percorsi ciclabili per un sistema di bike sharing (SBS). Tecniche di machine learning vengono utilizzate per caratterizzare i percorsi in un dato giorno. In questo articolo viene presentato un metodo per rilevare gli utenti 'più fedeli' rispetto ai tipi di itinerario e un modello di rete bayesiana che enfatizza il collegamento tra le condizioni in cui si sceglie l'itinerario e il tipo di itinerario.*

Key words: Bike sharing, Bike itinerary choice, Machine learning, Bayesian Network

Sabrina Tiziana Gaito

Department of Computer Science "Giovanni degli Antoni" and Data Science Research Center, Università degli Studi di Milano, via Celoria, 18, 20133 Milan, Italy, e-mail: sabrina.gaito@unimi.it

Giancarlo Manzi

Department of Economics, Management and Quantitative Methods and Data Science Research Center, Università degli Studi di Milano, via Conservatorio, 7, 20122 Milan, Italy, e-mail: giancarlo.manzi@unimi.it

Giorgio Saibene

Hamburg Business School, University of Hamburg, Moorweidenstraße 18, 20148 Hamburg, e-mail: giorgio.saibene@studium.uni-hamburg.de

Silvia Salini

Department of Economics, Management and Quantitative Methods and Data Science Research Center, Università degli Studi di Milano, via Conservatorio, 7, 20122 Milan, Italy, e-mail: silvia.salini@unimi.it

Matteo Zignani

Department of Computer Science "Giovanni degli Antoni", Università degli Studi di Milano, via Celoria, 18, 20133 Milan, Italy, e-mail: matteo.zignani@unimi.it

1 Introduction

Urban mobility is receiving increasing attention as it is considered one of the most important dimensions of the so-called *smart city* [1]. Recent developments in urban planning management have led BSSs to be a viable complement to traditional public transport systems. However, there are some important quandaries in organizing a successful BSS, for example when rightly predicting the behavior of users, and avoid an uneven distribution of the bikes across the city.

In this paper we are concerned with the optimization of fixed BSS, i.e. BSS with docking stations, and implement a decision framework to help policy makers to obtain an optimal prediction of cyclists itinerary choice in the case of a consolidated docking station-based BSS, the *BikeMi* system in Milan, Italy. It is operated by a private company having a service contract with the Municipality of Milan and receiving in return discounted advertisement spaces nearby the docking stations.

Obviously, all the actors in this kind of BSS (municipalities, private companies and citizens) can benefit from it if the service is well run. One of the most important obstacles to the well-functioning of the service lies in the relocation of bikes across crowded stations and the behavior of users. In this paper we address the former problem focusing on users' itinerary choices. In particular, daily closed users' paths are particularly important because they are those paths that identify the 'faithful' users, i.e. users that tendentiously use the bike similarly day by day. We analyze in details these path choices using machine learning techniques.

2 Data

Data were daily collected from the BSS *BikeMi* on each bike trip from June 2015 to May 2018 forming an overall data set containing initially 11,771,185 records. The data set was formed by two categories of data: a group of variables related to the bike sharing process and renting, and a group of variables related to the atmospheric conditions (including pollution status). In the first group of variables we had the details about the bike sharing process and renting transaction, including client and bike ID, the type of bike (traditional or e-bike), information about check-out and check-in (station number, station slot number, date and time of rent, etc.), rental time in minutes, distance covered in metres, amount of CO₂ avoided, calories consumed by cyclists, while meteorological variables (average daily temperature, average daily atmospheric pressure, precipitation condition - i.e. rain or no rain) as well as air condition indicators (daily average amount of PM₁₀, PM₂₅, NO₂ and CO₂) were included in the second group.

The results that we report below refer to a shorter period of time. In particular, we limited the investigation to the last available year - 2018 - for a total of 151 days, 50858 users and 1,457,609 transactions. Moreover, since we are interested in the set of 'faithful' users only, we pre-processed the data set by filtering out sporadic users. Specifically, we denote as faithful user a person who was active for at least

50 days (on average at least 10 days per month) only; we assume that the user was active on a specific day if he/she performed at least one transaction on that day. A further manipulation of the data set concerned the creation of the users' daily bike itineraries, since the originally released data set is centered on the single transaction enriched by identification and context variables. To this aim, we first reconstructed the transaction sequence of each client by leveraging the unique client ID field associated to each transaction. Operationally, we grouped the check-in and check-out transactions by the client ID and we ordered them chronologically. Then, to remove the circadian rhythm, we subdivided users' sequences into daily itinerary; thus, for each user, we obtain as many sequences as the number of days he/she was active. So, an itinerary travelled by a user u at day d is a sequence $\langle s_1, s_2, \dots, s_{n-1}, s_n \rangle$ of docking stations s_i , where even indexes indicate check-out stations and odd indexes denote check-in dock stations. Finally, we re-scaled the sequences so that they started from 0. For instance, the itinerary $\langle 10, 2, 2, 3, 67, 10 \rangle$ was re-scaled to $\langle 0, 1, 1, 2, 3, 0 \rangle$. In this way, we were able to highlight the typical behaviors of the users on riding, rather than focusing on the used docking stations. For this reason, we denoted the scaled itinerary as *itinerary type*. The entire process has been represented in Fig. 1 and resulted in a data set made up of 418,591 itineraries.

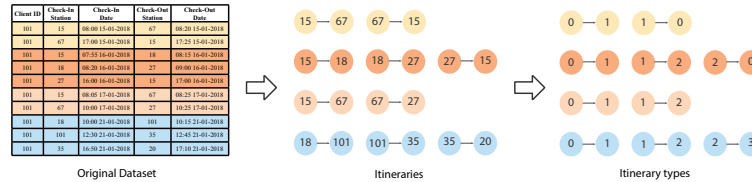


Fig. 1 Process for extracting daily itineraries of the 'faithful' users from the original data set.

We associated to each itinerary the context information inherited from the corresponding transactions, i.e. the first check-in and last check-out time slots (from 1 to 24), the weekday of the rent (from 1 to 7), the client ID, the average/minimum/maximum temperature of the day, the length and the duration of the whole itinerary, the number of visited dock stations, and the rain condition. Finally, we also introduced a boolean variable which indicates whether or not the user comes back to the first dock station of the itinerary. In the first case we indicate the itinerary as "closed", while in the latter the itinerary is "open".

3 Results

In this preliminary work we first tried to study the types of itineraries on the basis of the heterogeneity and the frequency with which they occur. Table 1 shows the corresponding Gini coefficients for the first 10 itineraries in terms of frequency. For each itinerary the Gini heterogeneity coefficient (calculated using the users as categories)

is obtained on the basis of the number of times each user chooses that itinerary at least once. It will be noted that heterogeneity is high; this means that the particular itinerary does not discriminate for selecting 'faithful' users. However, considering for each user the itineraries he/she chooses daily and considering a Gini heterogeneity (calculated using the itineraries as categories) threshold of 0.8 and choosing always the same itinerary type the number of users 'faithful' for a particular itinerary can be obtained (Table 2). So, for example, 2331 users can be considered preferring the itinerary type (0, 1), 455 users can be considered preferring the itinerary type (0, 1, 1, 0), and so on. Fig. 2 displays the CDF with respect to the Gini heterogeneity coefficient computed in this way and the red bar is the chosen threshold. Finally Fig. 3 shows the map of the first 20 docking stations from which most of open (red circles) and closed (blue circles) itineraries originate.

Table 1 The Gini heterogeneity index measured for the ten most common itinerary types.

Itinerary type	Gini Heterogeneity	Number of users
(0, 1)	0.999957	5603
(0, 1, 1, 0)	0.999813	5248
(0, 1, 1, 2)	0.999788	5237
(0, 1, 2, 3)	0.999798	4859
(0, 1, 2, 0)	0.999757	4817
(0, 1, 1, 2, 2, 3)	0.999551	1672
(0, 1, 1, 2, 3, 4)	0.999630	1494
(0, 1, 2, 3, 3, 4)	0.999583	1484
(0, 0)	0.998336	1426
(0, 1, 1, 2, 2, 0)	0.999004	1418

Table 2 Distinguishing itinerary types.

Itinerary Type	Number of users
(0, 1)	2231
(0, 1, 1, 0)	455
(0, 1, 1, 2)	119
(0, 1, 2, 0)	54
(0, 1, 2, 3)	31
(0, 0)	6
(0, 1, 1, 0, 0, 1, 1, 0)	1
(0, 1, 1, 2, 2, 1, 1, 0)	1

We also performed a Bayesian network (BN) analysis having the itinerary type as target variable. The question we wanted to address was the following: can the itinerary type, and perhaps its final station, be predicted knowing the weather conditions, the weekday, the starting station, the starting time, etc.? BNs implements a graphical model structure known as a directed acyclic graph (DAG), enabling an effective representation of the joint probability distribution (JPD) over a set of random variables. The structure of a DAG is defined by a set of nodes and a set of

Fig. 2 CDF of the normalized Gini impurity on the sequence types of each bikers. The red vertical line indicates the threshold used to identify bikers having a distinguishing itinerary type.

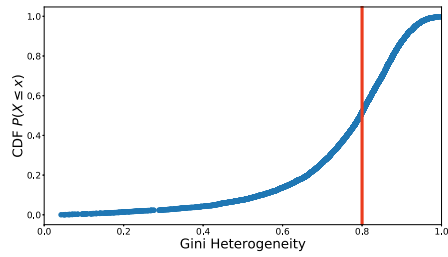
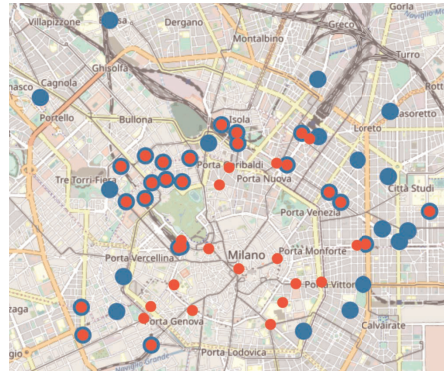


Fig. 3 Map displaying the first 20 docking stations from which most of open (red circles) and closed (blue circles) itineraries originate.



directed edges. The nodes represent random variables, whereas the edges represent direct dependencies among variables and are represented by arrows between nodes.

Fig. 4 shows a preliminary example of a BN applied to our data. A lot of arcs (edges) exist between sequence characteristics and target nodes. Using the JPD it is possible for example to predict the more probable sequence given the weather conditions (rain), the week day, the start section, the start hour, or to predict the more probable final station given the expected normalized sequence. One of the main benefit of BNs is that they provide an opportunity to conduct what if sensitivity scenarios.

4 Conclusion

This paper sought to evaluate the impact of closed itinerary choices of bike-sharing users in the overall functioning of the system. The high heterogeneity detected in the itinerary type analysis highlights the complexity of traffic forecasting. On the other hand, the first attempt to model JPD encourages the use of such models, that allow to simulate what-if sensitivity scenario. We limited the analysis to the last available year - 2018 - and only a subset of variables; further analysis and robustness tests are therefore needed. Future work will consider adding other data sources. Further information on users are available from the annual customer satisfaction survey.

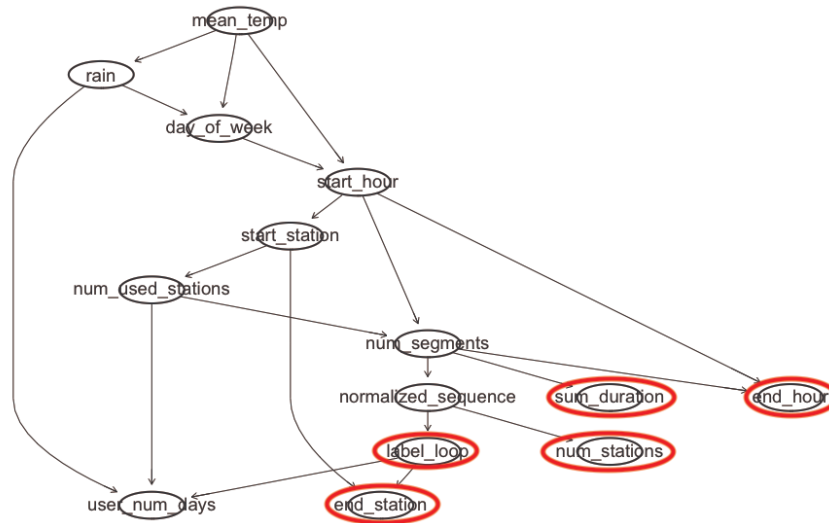


Fig. 4 BN obtained using score-based algorithms Hill-Climbing greedy search. Data-driven approach is used, no prior knowledge is imposed. Possible target nodes are highlighted in red.

We plan also to analyze neighboring stations together and the characteristics of the public transport network and of the commercial endowment of the area in which they are located. Other machine learning techniques, already applied in the BSS field [3] [4] will be taken into consideration.

References

- [1] Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., Meijers E. (2007). Smart cities. Ranking of European medium-sized cities, Final Report, Centre of Regional Science, Vienna UT. Available at: http://www.smart-cities.eu/download/smart_cities_final_report.pdf.
- [2] Lathia, N., Ahmed, S., Capra, L. (2012). Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation Research, Part C*, 22, pp. 88- 102.
- [3] Yang, Z., Hu, J., Shu, Y., Chang, P., Chen, J., Moscibroda, T. (2016). Mobility Modelling and Prediction in Bike-Sharing Systems. *Proceedings of the ACM International Conference on Mobile Systems, Applications, and Services (MobiSys 16)*, Singapore, pp. 165-178.
- [4] Manzi, G., Salini, S., Villa, C. (2019). Predicting Cycling Usage for Improving Bike-Sharing Systems. *Proceedings of the Second international conference on data science and social research (DSSR2019)*, Milan.

A PARAFAC-ALS variant for fitting large data sets

Una variante del PARAFAC-ALS per approssimare data set di grandi dimensioni

Michele Gallo, Violetta Simonacci and Massimo Guarino

Abstract The PARAFAC-ALS algorithm is the most widely used procedure for approximating arrays with a trilinear structure because it provides least squares solutions and delivers consistent outputs. Nonetheless, it is particularly slow at converging especially under challenging conditions, i.e. data multicollinearity, high factors' congruence and over-factoring. This shortcoming can be quite problematic when dealing with three-way arrays of large dimensions.

More efficient procedures can be employed, such as ATLD, however they are far less reliable. As an alternative, ATLD and ALS can be combined in a multi-optimization procedure in order to increase efficiency without reducing accuracy. This novel approach has been carried out and tested on artificial and real data.

Abstract *L'algoritmo PARAFAC-ALS è la procedura più utilizzata per approssimare array con struttura trilineare poiché fornisce soluzioni in termini di minimi quadrati ed è coerente nei suoi output. Tuttavia è particolarmente lento a convergere soprattutto in condizioni difficili come quelle di multicollinearità dei dati, elevata congruenza dei fattori e sovra-fattorizzazione. Questa carenza diventa piuttosto problematica nel trattare array di grandi dimensioni.*

È possibile impiegare altre procedure più efficienti, come l'ATLD, tuttavia sono generalmente meno affidabili. In alternativa, ATLD e ALS possono essere combinati in una procedura di ottimizzazione multipla al fine di aumentare l'efficienza senza ridurre la precisione. Questo nuovo approccio è testato su dati artificiali e reali.

Michele Gallo

Università degli Studi di Napoli - L'Orientale", DISUS, Largo S. Giovanni Maggiore, 30, Napoli,
e-mail: mgallo@unior.it

Violetta Simonacci

Università degli Studi di Napoli - L'Orientale", DISUS, Largo S. Giovanni Maggiore, 30, Napoli,
e-mail: vsimonacci@unior.it

Massimo Guarino

Università degli Studi di Napoli - L'Orientale", DISUS, Largo S. Giovanni Maggiore, 30, Napoli,
e-mail: mguarino@unior.it

Key words: ATLD, computational efficiency, CP model, trilinear data

1 Introduction

Trilinear tools such as the CANDECOMP/PARAFAC (CP) model [1, 5] are the most appropriate methods for obtaining a low-rank approximation of data arranged along three dimensions. By estimating three groups of parameters, one for each mode of the analysis, these techniques maintain the threefold structure of the data. Nevertheless, this gain in precision comes with a computational cost due to the high number of degrees of freedom.

The standard choice when estimating the CP model is the PARAFAC-ALS (ALS) algorithm, which has well-defined properties and ensures least-squares results. However, this procedure can be quite cumbersome, especially on data sets of large dimensions [6]. As a solution, alternative faster procedures have been proposed in the literature and compared to standard ALS in multiple studies [3, 8, 10, 11]. Two similar approaches stand out in terms of efficiency: ATLD (Alternating TriLinear Decomposition) proposed by [9] and SWATLD (Self-Weighted Alternating Tri-Linear Decomposition) introduced by [2]. They both use three loss functions rather than just one and tend to converge much faster than ALS. In addition, they perform exceptionally well in case of over-factoring and multicollinear data. Nonetheless, they are not consistent in providing acceptable final results and rarely yield least squares solutions. In general, ATLD is faster but far less stable than SWATLD.

Due to these limitations SWATLD and ATLD are rarely used in substitution of standard ALS. As an alternative, an integrated procedure called INT was proposed in [4] with the purpose of merging the benefits of SWATLD with those of ALS by concatenating their steps in a multi-optimization algorithm. This method was only tested for data that are particularly challenging for ALS because intrinsically multicollinear, i.e. compositional data. Results proved that INT was a good solution in that particular instance.

Following this approach, the next logical step is to verify the feasibility of initializing ALS with ATLD steps instead. Potentially, this methodology could be quite beneficial in terms of efficiency. Therefore, the purpose of this work is to implement a new ATLD-ALS integrated procedure (INT-2) and to assess its performance against ALS and INT.

2 Methodology

In this section the CP model and the PARAFAC-ALS algorithms are introduced. Successively an overview of the ATLD procedure and the suggested multi-optimization method INT-2 are presented.

2.1 The CP model with PARAFAC-ALS estimation

A tridimensional array $\underline{\mathbf{X}}$ contains values collected for I observations over a set of J variables in different K occasions. Such data can also be rearranged in three groups of two-way matrices, namely, K frontal slabs \mathbf{X}_k ($I \times J$), I horizontal slabs \mathbf{X}_i ($J \times K$) and J vertical slabs \mathbf{X}_j ($I \times K$).

If the latent information in $\underline{\mathbf{X}}$ is perfectly trilinear, the CP method is the best tool for extracting an approximate solution. This model is one of the generalizations of bilinear PCA to higher order arrays which aims to estimate three sets of parameters collected in the loading matrices \mathbf{A} ($I \times F$), \mathbf{B} ($J \times F$) and \mathbf{C} ($K \times F$), where F is the number of extracted loadings. The CP procedure can be written as follows:

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^t + \mathbf{E}_k \quad k = 1, \dots, K \quad (1)$$

$$\mathbf{X}_i = \mathbf{B}\mathbf{D}_i\mathbf{C}^t + \mathbf{E}_i \quad i = 1, \dots, I \quad (2)$$

$$\mathbf{X}_j = \mathbf{C}\mathbf{D}_j\mathbf{A}^t + \mathbf{E}_j \quad j = 1, \dots, J \quad (3)$$

Here \mathbf{E}_k ($I \times J$), \mathbf{E}_i ($J \times K$) and \mathbf{E}_j ($I \times K$) are the frontal, horizontal and vertical slabs of the array of residuals $\underline{\mathbf{E}}$. \mathbf{D}_k , \mathbf{D}_i and \mathbf{D}_j are diagonal matrix extracting the k th, i th and j th rows of \mathbf{C} , \mathbf{A} and \mathbf{B} respectively.

The CP model was originally fitted in a least squares sense [5] by estimating parameters using the ALS algorithm. The loss function of ALS aims to minimize the error term and can be written as:

$$\min_{\mathbf{A}, \mathbf{B}, \mathbf{C}} = \sum_{k=1}^K \|\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^t\|^2 \quad (4)$$

where $\|\cdot\|$ is the Frobenius norm.

This algorithm presents several important merits which include: a monotonically decreasing Loss of Fit; good performance in moderate noise scenarios; guaranteed convergence; and stable results. However, it is slow at converging and encounters additional difficulties in case of high collinearity, over-factoring and bad initialization.

2.2 Increasing efficiency: ATLD and INT-2

ATLD was proposed by [9] with the purpose of providing a more efficient estimating alternative which would also work better than ALS when over-factoring. In this perspective, three objective functions, one for each loading matrix, were assumed:

$$L(\mathbf{C}) = \sum_{k=1}^K \|\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^t\|^2, \quad (5)$$

$$L(\mathbf{A}) = \sum_{i=1}^I \|\mathbf{X}_i - \mathbf{B}\mathbf{D}_i\mathbf{C}^t\|^2, \quad (6)$$

$$L(\mathbf{B}) = \sum_{j=1}^J \|\mathbf{X}_j - \mathbf{C}\mathbf{D}_j\mathbf{A}^t\|^2. \quad (7)$$

Fastest convergence is guaranteed by the use of three loss functions with different response surfaces. However, by focusing only on the trilinear structure of the data, ATLD is not likely to render a solution in the least squares sense, it is more prone to degeneracy and may not monotonically decrease. Moreover model uniqueness requires the stricter condition that the three loading matrices must be full column rank.

Given these drawbacks, ATLD should not be used as a stand-alone procedure under undefined conditions. It was found, however, that concatenating ATLD optimization steps to ALS in an INTEgrated ATLD-ALS procedure (INT-2) could provide a relevant boost in efficiency without undermining all the advantages of ALS.

INT-2 is thus a composite procedure which is organized in two successive optimization stages. To start parameters are estimating through ATLD in order to obtain good initial estimates in a few iterations. In a second stage these preliminary ATLD outputs are improved and refined by carrying out ALS optimization to final convergence.

3 Simulation design and results

In order to test the effectiveness of the methodology proposed in this work, a simulation study is implemented. The performance of INT-2 is compared with standard ALS and with the alternative multi-stage approach INT in a Monte Carlo design. Large artificial data sets with varied levels of latent dimensionality, factor congruence and noise (homoscedastic and heteroscedastic) are generated and approximated with all considered algorithms. All procedures are performed by using the correct number of factors and by over-factoring.

When comparing results, the first step is to attest if INT-2 can guarantee stable enough results, similar to those of standard ALS and INT. This is checked by looking at the minimum value reached by the loss function and by calculating the percentages of successful identification of the correct solution. This condition is verified when the Tucker's congruence coefficient [7] between estimated and artificially generated factors is 0.9 or greater.

Once it is demonstrated that an acceptable level of accuracy is ensured, the increase in efficiency provided by INT-2 is assessed. Three efficiency diagnostics were taken into account: total CPU time employed to reach convergence; total number of iterations; occurrence of 2-factor degeneracies (2FDs).

2FDs, also referred to as temporary degeneracies, occur when the triple cosine (TC) becomes smaller than -0.8 . The TC among two generic factors f and f' is given by the multiplication of the three corresponding UCCs (uncorrected correlation coefficients) calculated for each loading matrix. For two generic factors of the first mode the UCC can be defined as:

$$\text{UCC}(\hat{\mathbf{a}}_f, \hat{\mathbf{a}}_{f'}) = \cos(\hat{\mathbf{a}}_f, \hat{\mathbf{a}}_{f'}) = \frac{\hat{\mathbf{a}}_f \cdot \hat{\mathbf{a}}_{f'}}{\sqrt{(\hat{\mathbf{a}}_f \cdot \hat{\mathbf{a}}_f)(\hat{\mathbf{a}}_{f'} \cdot \hat{\mathbf{a}}_{f'})}} \quad (8)$$

Preliminary results confirm that INT-2 is a fast, reliable option which outperforms INT in terms of efficiency without a significant loss in stability. Thus, it is particularly suitable for large data sets. Overall the procedure provides a reduction in time with respect to ALS larger than 50 %.

References

1. Carroll J. D., Chang J.J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. *Psychometrika* 35(3), 283–319 (1970)
2. Chen Z.P., Wu H.L., Jiang J.H., Li Y., Yu R.Q.: A novel trilinear decomposition algorithm for second-order linear calibration. *Chemometrics and Intelligent Laboratory Systems* 52(1):75–86 (2000)
3. Faber N.K.M., Bro R., Hopke P.K.: Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometrics and Intelligent Laboratory Systems* 65(1):119–137 (2003)
4. Gallo M., Simonacci V., Di Palma M.A.: An integrated algorithm for three-way compositional data. *Quality & Quantity*, 1–8 (2018)
5. Harshman R.A.: Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis (1970)
6. Kiers H.A.: A three-step algorithm for CANDECOMP/PARAFAC analysis of large data sets with multicollinearity. *Journal of Chemometrics: A Journal of the Chemometrics Society* 12(3):155–71 (1998)
7. Lorenzo-Seva U., Ten Berge J.M.: Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology* 2(2):57–64 (2006)
8. Tomasi G., Bro R.: A comparison of algorithms for fitting the PARAFAC model. *Computational Statistics & Data Analysis* 50(7):1700–1734 (2006)
9. Wu H.L., Shibukawa M., Oguma K.: An alternating trilinear decomposition algorithm with application to calibration of hplc–dad for simultaneous determination of overlapped chlorinated aromatic hydrocarbons. *Journal of Chemometrics* 12(1), 1–26 (1998)
10. Yu Y.J., Wu H.L., Nie J.F., Zhang S.R., Li S.F., Li Y.N., Zhu S.H., Yu R.Q.: A comparison of several trilinear second-order calibration algorithms. *Chemometrics and Intelligent Laboratory Systems* 106(1):93–107 (2011)
11. Yu Y. J., Wu H. L., Kang C., Wang Y., Zhao J., Li Y. N., Liu Y.J., Yu R. Q.: Algorithm combination strategy to obtain the second-order advantage: simultaneous determination of target analytes in plasma using three-dimensional fluorescence spectroscopy. *Journal of Chemometrics*, 26(5), 197–208 (2012)
12. Zhang S.R., Wu H.L., Yu R.Q.: A study on the differential strategy of some iterative trilinear decomposition algorithms: PARAFAC-ALS, ATLD, SWATLD, and APTLD. *Journal of Chemometrics* 29(3):179–192 (2015)

A Convex Mixture Model for Binomial Regression

Un modello mistura convessa per la Regressione Binomiale

Luisa Galtarossa and Antonio Canale

Abstract In this paper we introduce a convex mixture regression model to infer the effect of a potentially adverse exposure on a binary health outcome. Our construction—inspired by the recent contribution in [2]—assumes two extreme probabilities of observing a negative outcome at extreme doses, and relies on a convex combination of these extremal probabilities at each intermediate dose level. Inference is conducted by means of a Bayesian approach introducing a Gibbs sampler with closed-form full conditional posterior distributions.

Abstract *In questo articolo viene introdotto un modello mistura convessa per studiare l'effetto di un potenziale fattore di rischio su risposte biologiche dicotomiche. Tale formulazione—inspirata dal recente articolo in [2]—prevede la definizione di due probabilità per livelli di dose estremi e utilizza una combinazione convessa di tali probabilità per valori intermedi. L'inferenza è condotta attraverso un approccio bayesiano introducendo un campionamento di Gibbs con distribuzioni a posteriori condizionate in forma chiusa.*

Key words: Dose-response model, Bayesian regression, Extra risk function, Quantitative risk assessment.

1 Introduction

Our focus is to study how the probability distribution of a binary outcome $y \in \{0, 1\}$ varies with a predictor $x \in X \subseteq \mathbb{R}$. Specifically we are motivated by environmental applications in which the predictor is the dose of a potentially adverse exposure—

Luisa Galtarossa

Università degli Studi di Padova, Padova, Italy, e-mail: luisa.galtarossa@studenti.unipd.it

Antonio Canale

Università degli Studi di Padova, Padova, Italy, e-mail: canale@stat.unipd.it

such as an air or water pollutant—and the response is a binary health outcome—such as a premature delivery. There is a wide literature on dose–response modeling, relating dose to the risk of an adverse health outcome, with such models forming the basis of quantitative risk assessment and regulatory guidelines on safe levels of exposure. In quantitative risk assessment, one attempts to estimate the dose level associated with a selected small increase in risk relative to the background risk corresponding to zero dose. From a statistical perspective, it is then crucial to make robust and precise inferential considerations on some measure of risk. For binary responses, the proportional increase in risk at dose x is typically quantified via the *extra risk* function [3] defined as

$$r_E(x) = \frac{\Pr(y_i = 1|x_i = x) - \Pr(y_i = 1|x_i = 0)}{\Pr(y_i = 0|x_i = 0)} = \frac{\pi_x - \pi_0}{1 - \pi_0}, \quad (1)$$

where π_x denotes the probability of an adverse health outcome ($y_i = 1$) for a subject exposed to dose x .

Traditional dose-response models—such as logistic or Weibull [5, 6]—typically consider only decreasing or nondecreasing parametric functions while a more flexible nonparametric alternative may be able to characterise a variety of behaviours. In this paper we exploit a convex mixture regression recently introduced in [2] to model the distribution of continuous responses and adapt it to the binary responses case.

In the next section we introduce the model and comment on some of its properties. Section 3 introduces a Gibbs sampling algorithm for posterior inference under a Bayesian paradigm. Section 4 concludes the paper discussing future directions.

2 A Convex Mixture Model for Binomial Regression

Let $p_x(y)$ be the conditional probability mass function of y given $x \geq 0$ with

$$p_x(y) = p(y|x) = \pi_x^y (1 - \pi_x)^{1-y}.$$

We model the probability π_x as a convex combination of two extreme probabilities π_0 and π_∞ at $x = 0$ and $x \rightarrow \infty$ respectively, obtaining

$$\pi_x = \{1 - \beta(x)\}\pi_0 + \beta(x)\pi_\infty, \quad x \geq 0, \quad (2)$$

with $\beta(x)$ a monotone nondecreasing function that allows a continuous shifting in π_x from π_0 to π_∞ as x grows. Under this formulation it is important to define the extreme probabilities π_0 and π_∞ . Several choices can be made here. For example both the extreme probabilities can be unknown and are estimated from the data. Another opposite possibility is to fix both parameters and set $\pi_0 = 0$ and $\pi_\infty = 1$. In this paper, we mediate between these two possibilities and assume that π_0 is unknown while $\pi_\infty = 1$ is fixed. This assumption is consistent with the idea that a

negative outcome (e.g. death, tumour, premature delivery) happens with probability one at infinite dosage of the adverse exposure and that at no dose there may be other (not observed) factors that lead to the adverse outcome.

To complete the specification of the model we need to define a flexible function $\beta(x)$. Consistently with the hypothesis that increasing levels of exposure lead to increasing risk, it is important to define this function under the condition of monotonicity. To this end we follow the definition presented in [2] assuming

$$\beta(x) = \sum_{j=1}^J w_j \psi_j(x), \quad x \geq 0, \quad (3)$$

where $\psi_1(x), \dots, \psi_J(x)$ are monotone nondecreasing functions with $\psi_j(x) \in [0, 1]$ for all $j = 1, \dots, J$ and $\sum_{j=1}^J w_j = 1$. The default choice for $\psi_1(x), \dots, \psi_J(x)$ is the I-splines basis [4]. The effect of the dose x enters only via the flexible function $\beta(x)$. It is important to notice that while $\beta(x)$ interpolates the two extremal probabilities π_0 and π_∞ , we cannot observe outcomes related to π_∞ , as $x \rightarrow \infty$ is never observed in practice.

An appealing consequence of our formulation is that we can directly interpretate the *extra risk* function in terms $\beta(x)$. Indeed, under this formulation the *extra risk* function (1) turns out to be

$$r_E(x) = \frac{\pi_x - \pi_0}{1 - \pi_0} = \frac{\{1 - \beta(x)\}\pi_0 + \beta(x)\pi_\infty - \pi_0}{1 - \pi_0} = \frac{(1 - \pi_0)\beta(x)}{1 - \pi_0} = \beta(x). \quad (4)$$

3 Bayesian Inference

Following a Bayesian approach we elicit suitable prior distributions for all the unknown model's parameters.

For the coefficients of the basis expansion of $\beta(x)$ in (3) we define a Dirichlet prior so that $w = (w_1, \dots, w_J)^T \sim \text{Dir}(\eta)$ where $\eta = (\eta_1, \dots, \eta_J)$ is the vector of hyperparameters.

For π_0 we assume a beta distribution, $\pi_0 \sim \text{Beta}(a, b)$ with $a, b \geq 0$. The parameters a and b need to be elicited possibly using prior information about the problem at hand. For example, when y is a binary variable indicating a pre-term/term pregnancy, a and b can be set at 27 and 360, respectively, so that the prior expectation of π_0 is roughly 0.07—i.e. equal to the proportion of preterm births in developed countries as showed in [1]—and the probability of obtaining $\pi_0 \leq 0.05$ —representing the inferior preterm birth rate presented in [1]—is very low, i.e. $\Pr(\pi_0 < 0.05) = 0.05$.

Before describing the Gibbs sampling algorithm we introduce augmented variables (b_i, d_i) , for each $i = 1, \dots, n$. Consistently with the formulation described in the previous section, $b_i \in \{1, \dots, J\}$ indicates the basis function associated with unit i and $d_i \in \{0, 1\}$ represents an indicator of the membership to one of the two extreme binomial distributions. Conditioned on this augmented variables it is possible

to define the following relations:

$$\{\pi_{x_i}|d_i\} = (1 - d_i)\pi_0 + d_i, \quad d_i \sim \text{Bern}\{\beta(x_i)\}, \quad (5)$$

$$\{\beta(x_i)|b_i = j\} = \psi_j(x_i), \quad b_i \sim \text{Cat}(w_1, \dots, w_J), \quad (6)$$

where $\text{Cat}(w_1, \dots, w_J)$ denotes a categorical variable with probabilities (w_1, \dots, w_J) and $\text{Bern}\{\beta(x_i)\}$ denotes a Bernoulli distribution with success probability $\beta(x_i)$.

Samples from the joint posterior distribution of the parameters can be obtained by sampling iteratively from the following steps.

1. Update b_i from the full conditional categorical random variable having probabilities

$$\Pr(b_i = j|-) \propto w_j[\{1 - \psi_j(x_i)\}p_0(y_i) + \psi_j(x_i)p_\infty(y_i)],$$

for every $j = 1, \dots, J$.

2. Update w from the full conditional Dirichlet distribution

$$(w|-) \sim \text{Dir}(\eta_1 + n_1, \dots, \eta_J + n_J),$$

where n_j is the number of subject in which $b_i = j$, and update $\beta(x)$ by applying equation (3).

3. Update d_i from

$$(d_i|-) \sim \text{Bern}\left[\frac{\beta(x_i)p_\infty(y_i)}{\{1 - \beta(x_i)\}p_0(y_i) + \beta(x_i)p_\infty(y_i)}\right].$$

4. Finally update π_0 , from

$$(\pi_0|-) \sim \text{Beta}\left(a + \sum_{i:d_i=0} y_i, b + n_0 - \sum_{i:d_i=0} y_i\right)$$

where n_0 is the number of units with $d_i = 0$.

4 Conclusion

In this paper we presented a modification of the convex mixture regression presented in [2] dealing with binary health outcome. The model is particularly interesting in quantitative risk assessment applications as we showed that the $\beta(x)$ function can be directly interpreted as *extra risk* function. A detailed simulation study is subject to ongoing investigation to assess the empirical performance of the proposed method in a variety of situations.

Acknowledgement

The authors are supported by the University of Padova under the STARS Grants programme BNP-CD.

References

- [1] S. Beck, D. Wojdyla, L. Say, A. P. Betran, M. Merialdi, J. H. Requejo, C. Rubens, R. Menon, and P. F. Van Look. The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity. *Bulletin of the World Health Organization*, 88(1), 2010.
- [2] A. Canale, D. Durante, and D. B. Dunson. Convex mixture regression for quantitative risk assessment. *Biometrics*, 2018.
- [3] A. F. Filipsson, S. Sand, J. Nilsson, and K. Victorin. The benchmark dose method-review of available models, and recommendations for application in health risk assessment. *Critical Reviews in Toxicology*, 33(5):505–542, 2003.
- [4] J. O. Ramsay. Monotone regression splines in action. *Statistical Science*, 3:425–441, 1988.
- [5] C. Ritz, F. Baty, J. C. Streibig, and D. Gerhard. Dose-response analysis using r. *PloS one*, 10(12):e0146021, 2015.
- [6] C. Ritz and J. C. Streibig. Bioassay analysis using R. *Journal of Statistical Software*, 12(5):1–22, 2005.

Blockchain as a universal tool for business improvement

Massimiliano Giacalone, Diego Carmine Sinitò, Emilio Massa, Federica Oddo, Enrico Medda, Vito Santarcangelo

Abstract: The aim of this work is to present the characteristics of the blockchain technology and its potential in corporate case study applications. The paper presents in detail an example of the implementation of permissioned blockchain and other examples of blockchain (also of semantic type) applied to the temporal certification of business processes of some brilliant southern Italy realities.

Key words: blockchain, bitcoin, safety for business

1 What is the Blockchain and how it works

Thanks to the fame of the bitcoin cryptocurrency, lately also blockchain technology has become object of great attention. Blockchains, thanks to their properties, are becoming an increasingly popular tool for companies that need to certify and keep their data safe.

1.1 *Blockchain and Bitcoin*

If we would ask someone what a blockchain is, they probably would link the answer to the word bitcoin. But how are bitcoin and blockchain related? These two entities were born simultaneously, however the blockchain is independent from bitcoin and the demonstration of this is in the evolution and the birth of new types of blockchain. Satoshi Nakamoto, pseudonym of the creator of cryptocurrency bitcoin, in his paper "Bitcoin: A Peer-to-Peer Electronic Cash System" introduces a system for the exchange of a virtual currency that does not require a financial institution that certifies the transactions. In this new scenario the problem of double spending was solved by using a distributed register where all the transactions are kept, which bases its security on a system of signing transactions with public key - private key, hashes and a proof encryption. The register obtained would be immutable unless after an

alteration the attacker had sufficient computing power to perform the cryptographic test again for all the blocks that make up the register before the other participants in the network.

1.2 *Blockchain: structure and types*

When a user makes a transaction on the bitcoin network, this is announced publicly to all the nodes that are part of it, in this way it is possible to certify that the transaction took place in a specific time (timestamp). Each node collects all new transactions by forming a block. When a block is complete, it can take part of the chain after passing a cryptographic test. The proof-of-work at the base of bitcoin is the increment of a numerical value, called *nonce*, which causes the block hash to start with a certain number of zeros. Once the cryptographic problem is solved, the node communicates to all the other nodes the solution found. Nodes accept blocking only if all transactions within are valid and they express their acceptance by creating the next block which includes the hash of the block that has just been accepted. Thanks to its distributed structure and the large number of nodes, the blockchain guarantees the immutability of data and business continuity.

This kind of network is also called public, this means that anyone with a device connected to the internet can become a network node and can take part at the creation of blocks and the validation of new blocks or have access to the entire transaction register. Various types of blockchain which provide for a form of centralized authority have been recently developed. In permissioned blockchain, unlike public blockchain, access is restricted to some users only. Furthermore, the central authority defines the role of a user within the network and to which information he has access. There are different levels of access to the network, each with different functionalities:

- Reading the ledger, which may be subject to certain constraints or may be accessible for all nodes.
- The possibility of making transactions within the network, which must then be validated and inserted into the blockchain.
- Perform block mining and validation operations.

This more performing kind of blockchain, is therefore preferred by companies that want to maintain a high level of confidentiality of their data while benefiting from all the characteristics of the blockchain.

2 Safety for business

The blockchain is designed to keep safe and unaltered the data inside. The data are inserted into the blocks in way to create a chain held together by the integrity of the same data. If a single data is altered the whole chain is invalidated.

2.1 *Block composition*

This security is guaranteed by a particular string of 64 characters called *hash* which represents all the information encrypted within a block and must meet previously agreed criteria. This *hash* is an alphanumeric string of characters generated by the SHA-256 function. The block is made up of different fields based on the use that you want to make of the blockchain and the data you want to protect. There are some fields of the block that are essentials for the blockchain to be well structured and capable of being efficiently constructed and these fields are:

An **index**, which allows to uniquely identify each block; the **timestamp**, which allows to identify the exact moment in which the block is generated; the **hash** of the previous block. The last data allows to create the famous “chain”, each hash of a block will become a fundamental component of the next one. *Nonce* is a really important value for the respect of all the above mentioned criteria and will be further analyzed later; to all this basic information we must add all the ones we want to actually store safe and intact. For the realization of a new block, it was achieved a function which allows to obtain the index of the last block inserted.

2.2 *“Chain” creation*

It was created a function that permit, after obtaining this information, all of this data to be acquired. This function (*push ()*) takes care of acquiring information such as index, timestamp and all the data we want to preserve. Within this function, once the hash of the previous block has been obtained (through its index), the block index is incremented so as to become the index of the block that we are currently inserting. Next, the heart of the insertion: the function that deals with “undermining” and creating a hash that allows to respect the criterion.

This criterion is useful because it allows us to create a non-trivial and independent hash of the data we actually want to memorize, for this task is used the *nonce*. The *nonce* is initially set to zero. The hash is generated with all the data, index, timestamp, information to be stored, previous block hash and the *nonce*, then it is consulted a function that generates the hash through SHA-256 encryption. If the criterion is respected, the block is added to the chain, otherwise it enters in a *while* loop. The while condition is repeated until the hash is composed in such a way as to meet the previously decided requirement. For each iteration the nonce is

incremented by one, this allows to generate a completely different hash for each iteration, up to the requirement. The more difficult is the requirement to achieve, more iterations will be needed, greater will be potentially the nonce value. Once the wanted hash is obtained, the block is completed and the information obtained can be safely stored.

2.3 *Validation of the process*

All the information are stored in a database. For this reason there is the possibility that somehow they can be modified, so it is necessary to use an algorithm that allows to understand if the blockchain has been compromised and is invalidated. This is made possible thanks to a function that runs through the entire blockchain and recalculates the hashes of all the blocks: through a *for* loop, a specially created matrix crosses each block and verifies its integrity. At each iteration through the data present in the array, the hash of the block is recalculated and compared to the stored hash. The hash of the previous block must coincide with the previous hash of the current block, if one of these two conditions is not respected then the blockchain has been invalidated. If successful, the iteration will be repeated until the whole blockchain is verified.

3 Cases study analysis

We will show below several case studies in which the blockchain was used with various kind of implementation. This will explain the potential, the possible uses and developments of this new technology that holds great potential for all the multiplicity of services or processes which need information integrity, time validations and business continuity.

3.1 *L'Antincendio Srl*

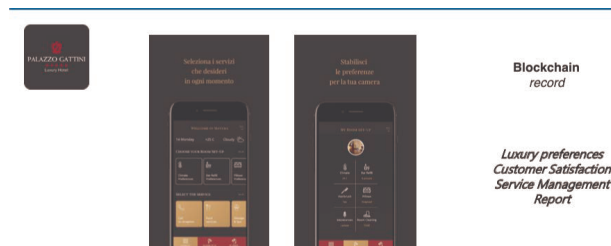
When we talk about services or products designed to guarantee people safety, the temporal certification and the integrity of the information about them becomes a fundamental factor. In this context the company L'Antincendio Srl has found an answer in the blockchain technology that has been used to map and certify all the maintenance and customer reports on each individual fire extinguisher. In this way you will not only have the possibility to reconstruct the time-certified history of each individual fire extinguisher, but you can also analyze the information in order to obtain the security level of the product with the certainty of data integrity. This is a clear example of how this technology has been used by a company to improve the level of safety of its products.

3.2 *Capurso Azienda casearia Srl*

The case study about Capurso Azienda casearia Srl is different, in this example the company, in order to protect the quality of its product and remain compliant with the DOP hygiene standards also in key 231, required an effective tool for mapping all the processes concerning the production of the raw material (cow's milk), in this case the blockchain was used to guarantee the integrity of the data during the audit of the entire production process of the raw material and to certify the history of every single dairy cow allowing the company to pursue increasingly higher quality standards.

3.3 *Palazzo Gattini*

A very special case study concerns the blockchain implemented for the Palazzo Gattini structure, which offers luxury services with high quality standards and interfaces with a very demanding clientele. In this case, it was decided to map not only the entire process concerning the service offer, but also the preferences of each individual user (luxury preference) and the customer satisfaction within the ledger. Given the strong semantic link between the customer and the information saved on the ledger, it has been used a semantic blockchain, that is to say that within each individual block there will also be the hash of the previous block concerning the client in question, in this way we can analyze in a better way the information collected to obtain the level of luxury compliance of the structure.



3.4 *Erreffe Srl*

We can find another example of the use of blockchain in the project manager of Erreffe srl, in this case all the tasks within the company have been mapped in the ledger with the related processes according to the individual tasks, it has been fundamental to certify all the progress in order of time in way to be able to view and analyze the evolution and of the progress with complete and certified data.

3.5 *Caldarola Srl*

In the last case study the blockchain technology is used by the company Caldarola Srl for the vehicle rental software. Here, all the information (generic or about maintenance) related to the individual vehicles present in the company are saved in the blockchain including picking and releasing. In this case the time certification has a great value in order to allow the company to know the current status of all vehicles and to be able to view the chronology of all the maintenances working on integral data and with certified time values without using particularly complex processes for this kind of certification.

4 Conclusions and further developments

This paper presented interesting application developments of permissioned blockchain applied to practical case studies of small companies in southern Italy. Incorporating the potential of this technology in the processes of southern SMEs is an indication of how much this technology is transversal and of great importance for business improvement and for meeting regulatory compliance requirements.

An example of application may concern the certification of machinery 4.0, as the interconnection of the same can be proved by blockchain recording of the data of the interconnection of individual machines, which would have probative validity. From this point of view, the process of processing certified log via blockchain, which would allow to compare the certified "real process" with the company procedures in order to determine a certified "process gap", can also be applied.

References

1. Nakamoto, S. (2009). Bitcoin: A Peer-to-Peer Electronic Cash System. <https://bitcoin.org/bitcoin.pdf>.
2. M.Colucci, UIBM, Patent, 27/12/2018 - 102018000021061, Sistema ad alta innovazione per la certificazione dei processi di manutenzione antincendio
3. F.Fanari, V.Santarcangelo,D.C. Sinitò, Esperienze di Ricerca e Sviluppo applicate alle brillanti realtà del nostro sud,RCE Multimedia, 2018
4. A. Brandonisio, UIBM, 28/12/2018 n. 102018000021313, Sistema basato su blockchain per la certificazione di filiera di prodotti caseari
5. E. Grassano, UIBM, 02/01/2019 rif n. 102019000000025, Sistema blockchain oriented per la certificazione dei servizi di una struttura luxury
6. R. Festa, C. Dell'Acqua, S. Burgi, UIBM, 2019, "Sistema intelligente per la gestione di commessa mediante blockchain in ottica BIM"
7. M.Masi, UIBM, Rif. 102018000007495, Sistema basato su blockchain per il controllo della flotta in compliance
8. M.Masi, SIAE 21/08/2018,Rif. 012763, Sagest Fleet Gdpr Blockchain

Seasonality in tourist flows: a decomposition of the change in seasonal concentration

La stagionalità nei flussi turistici: una scomposizione della variazione nella concentrazione stagionale

Luigi Grossi and Mauro Mussini

Abstract The change over time in seasonal concentration of tourist flows is broken down into two components. One component measures the contribution of the changes in the seasonal pattern of tourist flows. The second component measures the contribution of the change in the magnitude of seasonality.

Abstract *La variazione nella concentrazione stagionale dei flussi turistici è scomposta in due componenti. La prima componente misura il contributo dei cambiamenti nell'andamento stagionale. La seconda componente misura il contributo della variazione nell'intensità della stagionalità.*

Key words: decomposition, Gini index, seasonal concentration, tourist flows

1 Introduction

Tourist flows in a tourism destination are often characterized by fluctuations during the year. If these fluctuations recur with similar timing and magnitude from year to year, they are called seasonal fluctuations and are a systematic component of a tourism time series, namely seasonality [2]. Seasonality implies that tourist flows tend to be concentrated in certain periods of the year (peak periods), generating an unequal distribution of tourists among the periods [1]. From this perspective, the

¹ Luigi Grossi, Dipartimento di Scienze Economiche, Università degli Studi di Verona; email: luigi.grossi@univr.it

Mauro Mussini, Dipartimento di Economia, Metodi Quantitativi e Strategie di Impresa, Università degli Studi di Milano Bicocca; email: mauro.mussini1@unimib.it

degree of seasonality can be measured in terms of seasonal concentration by using a conventional concentration index, such as the Gini index or the Theil index.

Concentration indexes are descriptive statistical tools which were originally introduced in the literature on income inequality, but their field of application crossed that of the measurement of income inequality. Such indexes were used to quantify the level of seasonality in several studies examining the evolution of seasonality in tourist flows [3], [4], [5], [8]. The change over the years in seasonal concentration can be measured by observing the change in the concentration index. However, a basic element of seasonality cannot be monitored by examining solely the change in the concentration index. An essential feature of seasonality is that the intra-year fluctuations occur with the same timing from year to year. However, the value of a concentration index does not vary if fluctuations have the same magnitude but a different timing. Since a conventional concentration index is invariant to changes in the seasonal pattern, a measure of seasonal stability is needed in addition to a measure of the change in seasonal magnitude when analysing the evolution of seasonality. In this paper, we show that these measures can be obtained by decomposing the change in seasonal concentration. In particular, seasonal concentration in a given year is measured by using the Gini index [6], a very popular measure of concentration. After measuring seasonal concentration in two different years, the change in seasonal concentration is split into two components. One component is a measure of seasonal stability, which tracks changes in the seasonal pattern. The second component measures the change in seasonal magnitude, assuming that the seasonal pattern is stable.

2 Decomposing the change in seasonal concentration

Let Y stand for a variable representing either the frequency of tourist arrivals or the frequency of nights per stay in a destination. Let $y_{1,t}, \dots, y_{n,t}$ be the time series of tourists visiting a destination in year t , where n is the number of periods in a year (e.g., $n = 12$ for monthly observations) and $y_{i,t}$ is the tourist frequency of period i in year t . The Gini index measuring the concentration of tourists in t is

$$G_t = \frac{2cov[y_{i,t}, r(y_{i,t})]}{n\bar{y}_t}, \quad (1)$$

where \bar{y}_t is the average tourist frequency per period and $r(y_{i,t})$ is the rank of period i according to the increasing order of tourist frequency.

Now suppose that the monthly distribution of tourists is observed in year $t + k$. The change in the concentration of tourists from year t to year $t + k$ is measured by the difference between the Gini index in $t + k$ and the Gini index in t :

$$\Delta G = G_{t+k} - G_t = \frac{2cov[y_{i,t+k}, r(y_{i,t+k})]}{n\bar{y}_{t+k}} - \frac{2cov[y_{i,t}, r(y_{i,t})]}{n\bar{y}_t}. \quad (2)$$

A decomposition of the change in seasonal concentration

However, ΔG does not necessarily capture all the aspects of the change in seasonality. For instance, consider the numerical illustrations in figure 1 showing two monthly distributions in years t and $t + k$. The twelve absolute figures in $t + k$ are equal to those in t but the ranking of months by tourist frequency has changed. August and July have exchanged their frequencies and the same has been done by September and June, while the frequencies of the other months in $t + k$ are the same as those in t . In such a situation, the Gini index in $t + k$ is equal to the Gini index in t . ΔG equals zero, suggesting that seasonality has remained unchanged since t , but the distribution of tourists among months has changed. To capture the changes in seasonality depicted in figure 1, a measure of seasonal stability and a “pure” measure of the change in seasonal magnitude are needed. These measures are obtained by decomposing ΔG .

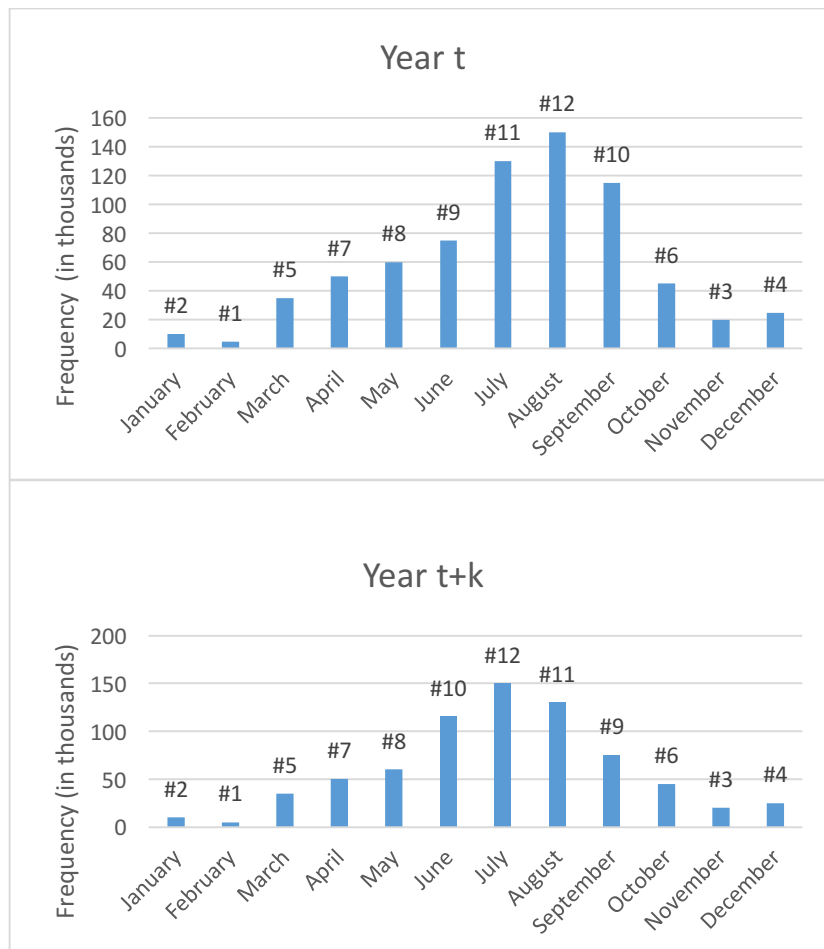


Figure 1: Monthly distributions of tourists in years t and $t + k$.

Let $C_{t+k|t}$ stand for the concentration coefficient of tourist frequencies in $t + k$ obtained by sorting periods according to their ranking in t instead of that in $t + k$:

$$C_{t+k|t} = \frac{2cov[y_{i,t+k}, r(y_{i,t})]}{n\bar{y}_{t+k}}. \quad (3)$$

By adding and subtracting $C_{t+k|t}$ to the right-hand side of equation (2), the difference in the Gini index is broken down into two components:

$$\Delta G = \left\{ \frac{2cov[y_{i,t+k}, r(y_{i,t+k})]}{n\bar{y}_{t+k}} - \frac{2cov[y_{i,t+k}, r(y_{i,t})]}{n\bar{y}_{t+k}} \right\} + \left\{ \frac{2cov[y_{i,t+k}, r(y_{i,t})]}{n\bar{y}_{t+k}} - \frac{2cov[y_{i,t}, r(y_{i,t})]}{n\bar{y}_t} \right\}$$

$$\Delta G = R + M. \quad (4)$$

R in equation (4) is the re-ranking component measuring the change in concentration due to the position exchanges between periods in the ranking of periods from t to $t + k$. The re-ranking component is analogous to the re-ranking measure used for capturing the re-ranking of income receivers when decomposing the change in income inequality [7]. R equals zero if the ranking of periods in $t + k$ is the same as that in t . The component R reaches its highest value, that is $2G_{t+k}$, when the ranking of periods in $t + k$ is completely reversed with respect to the ranking of periods in t . Since the re-ranking component captures any deviation in the seasonal pattern, R is seen as a measure of seasonal stability. The larger the re-ranking component, the more unstable the seasonal pattern.

The term M in equation (4) is the magnitude component measuring how much tourist frequencies in year $t + k$ are more (or less) concentrated on the periods with the highest ranks in year t . Since M is calculated by holding periods sorted according to their ranking in t , the seasonal pattern is supposed to be the same in t and $t + k$. The magnitude component is positive (negative) when the magnitude of seasonality has increased (decreased) over time. M is equal to zero if the tourist frequency distribution among periods has not changed or if the tourist frequencies of all periods have changed in same proportion, leaving unchanged the relative disparities between tourist frequencies.

3 Application

The decomposition is used to analyse the evolution of seasonality in tourist arrivals in a mature tourism destination of one of the most tourism-oriented Italian regions: the Veneto region. The tourism destination we consider is Jesolo-Eraclea, a popular Venetian seaside destination with a tourism product mainly based on sun and beach. Since sun and beach play a major role in the Jesolo-Eraclea tourism product, the distribution of tourist arrivals shows sharp peaks in the summer months and a high level of seasonal concentration. Data on tourist arrivals are from the database of the Veneto Region, which disseminates the monthly series of tourist arrivals in the Local Tourism Systems of the Veneto Region.

A decomposition of the change in seasonal concentration

The change in seasonal concentration over a decade is examined, setting 2006 and 2016 as the initial and final reference years respectively. A ten-year long period is considered in our analysis since we suppose it is a reasonable time interval to track changes in seasonality. The monthly distributions of tourist arrivals in Jesolo-Eraclea in 2006 and 2016 are displayed in figure 2. The seasonal concentration has increased from 0.437106 to 0.475891 between 2006 and 2016. The change in seasonal concentration $\Delta G = 0.038785$ is due to both an increase in seasonal magnitude, measured by $M = 0.020301$, and some changes in the seasonal pattern, measured by $R = 0.018484$. Figure 2 shows that tourist arrivals in July and August overtook those in June in 2016, whereas June was the month with the highest number of arrivals in 2006. These months exchanged their positions in the ranking of months by tourist arrivals, indicating the seasonality is not entirely stable. The contribution of the re-ranking of months reinforces that of the increase in the magnitude of relative disparities between months in terms of tourist arrivals, resulting in an increase of concentration of 0.038785.

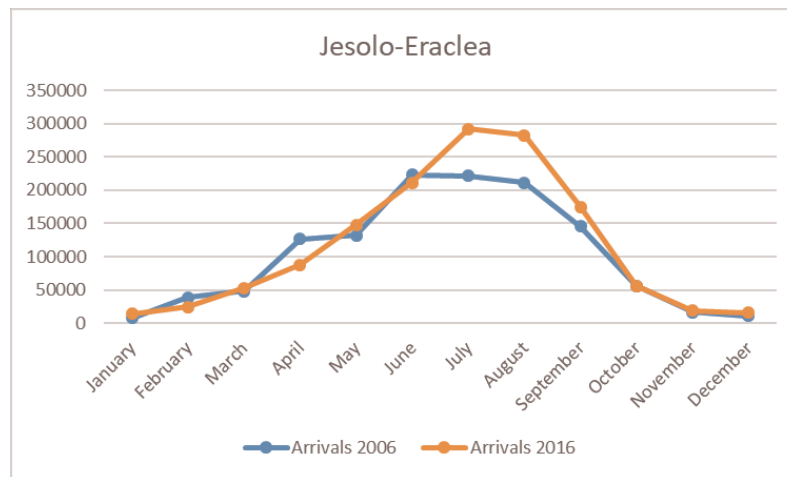


Figure 2: Monthly tourist arrivals in Jesolo-Eraclea in 2006 and 2016. Data were downloaded from the Veneto Region database in October 2018.

4 Conclusion

A common approach to measuring seasonal concentration in tourist flows is the application of a conventional concentration index to the intra-year distribution of tourist arrivals or nights spent, depending on the indicator expressing tourist flow. However, when measuring the change in seasonal concentration over time, observing the variation in the seasonal concentration index may lead to misleading conclusions

since a small change in the index can be the outcome of the offsetting contributions of changes in the seasonal pattern and magnitude. The decomposition of the change in seasonal concentration sheds light on the different aspects of a change in the seasonality of tourist flows, revealing the roles of changes in the seasonal pattern and magnitude. This decomposition provides local governments of tourism destinations with additional information on the change in seasonality of tourist flows. For instance, local policy makers can assess whether a counter-seasonal initiative has been successful in mitigating seasonality by looking at the seasonal magnitude component, while the seasonal stability component tells them whether a shift in the timing of fluctuations in tourist flows has occurred.

Natural directions for future research include the application of the decomposition to data on other tourism destinations and the development of a statistical procedure for testing hypotheses concerning the seasonal stability and magnitude components.

References

1. Allcock, J.: Seasonality. In: S. Witt, L. Moutinho (Eds.), *Tourism Marketing and Management Handbook* (pp. 387-392). London: Prentice Hall (1989)
2. Bar-On, R.V.: The Measurement of Seasonality and its Economic Impacts. *Tourism Economics*, 5, 437-458 (1999)
3. Fernández-Morales, A.: Decomposing seasonal concentration. *Annals of Tourism Research*, 30, 942-956 (2003)
4. Fernández-Morales, A., Mayorga-Toledano, M.C.: Seasonal concentration of the hotel demand in Costa del Sol: A decomposition by nationalities. *Tourism Management*, 29, 940-949 (2008)
5. Fernández-Morales, A., Cisneros-Martínez, J.D., McCabe, S.: Seasonal concentration of tourism demand: Decomposition analysis and marketing implications. *Tourism Management*, 56, 172-190 (2016)
6. Gini, C.: *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. Bologna: Tipografia di P. Cuppini (1912)
7. Jenkins, S., Van Kerm, P.: Trends in income inequality, pro-poor income growth, and income mobility. *Oxford Economic Papers*, 58, 531-548 (2006)
8. Lau, P., Koo, T.T.R., Dwyer, L.: Metrics to measure the geographic characteristics of tourism markets: An integrated approach based on Gini index decomposition. *Tourism Management*, 59, 171-181 (2017)

Are Real World Data the smart way of doing Health Analytics?

Real World Data: la base di una nuova ricerca clinica?

Francesca Ieva

Abstract Real world data (RWD) and real world evidence (RWE) are playing an increasing role in health care. Despite their use may strongly improve healthcare research as well as health related decision making, management and planning, many barriers remain to their use in clinical practice. The aim of this paper is to discuss some issues related to RWD aiming at proposing a new paradigm of healthcare research based on RWE.

Abstract *I dati dal mondo reale e le relative evidenze scientifiche stanno avendo un ruolo di sempre maggiore importanza nello scenario odierno della ricerca in ambito clinico. Nonostante la diffusione di tali dati possa comportare significativi miglioramenti nella ricerca sanitaria e nelle politiche decisionali e di gestione nell'ambito della salute, ancora oggi permangono molte barriere alla diffusione e al consolidamento del loro utilizzo nella prassi clinica. Lo scopo di questo lavoro discutere alcune questioni relative ai dati del mondo reale al fine di proporre un nuovo paradigma di ricerca nell'ambito della salute, basato sulle evidenze generate dai dati del mondo reale.*

Key words: Real World Data, Real World Evidence, Health Analytics

1 Background and setting

The terms “Real World Data” (RWD) and “Real World Evidence” (RWE) refer to the widely heterogeneous amount of data arising from current practice in any area, and to the evidence that may be pointed out applying suitable statistical methods to their analysis. RWD are playing an increasing role in health care research nowadays. The health care community is using RWD to support coverage decisions and to

Francesca Ieva

MOX - Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, via Bonardi 9, 20133, Milano (IT), e-mail: francesca.ieva@polimi.it

develop guidelines and decision support tools for use in clinical practice, to monitor postmarket safety and adverse events and to make regulatory decisions. Medical product developers are using RWD and RWE to support clinical trial designs and observational studies to generate innovative, new treatment approaches.

In other words, healthcare is rapidly transitioning to a new world of patient choice with a focus on outcomes and value, being value

$$VALUE = \frac{CLINICAL BENEFITS}{DIRECT COSTS}$$

Therefore, suitable indicators for “value” and suitable systems for monitoring processes, utilizations and performances in healthcare are more and more needed. Moreover, the use of computers, mobile devices, wearables, and other biosensors to gather and store huge amounts of health-related data has been rapidly accelerating.

Starting from these considerations, the “Real World” term and related data have to be properly defined, in order to give birth to a new definition of clinical evidence, that is real world based. The aim of this paper is to point out a discussion on what RWD are in the actual context, and how they may lead to a convincing Real World Evidence, in order to set a new paradigm for a smart research in clinical biostatistics.

2 Real World Data in practice

There is no univocal definition of what RWD are in practice. It is pretty well known the potential of RWD, especially into the healthcare setting, as well as the issues related to their use. Nevertheless, a clear definition of what RWD are has still to come. This is mainly due to the quick technological evolution which creates new type of data and new potential in data collection day by day.

In general, RWD are the data relating to patient health status and health consumption, as well as the delivery of health care routinely collected from a variety of sources. Among others, the main sources are: Electronic health records (EHRs), claims and billing activities, clinical registries, patient-generated data and data gathered from external sources that can inform on health status, such as mobile devices.

For these reasons, RWD sources generally fall into four categories, which are likely to be expanded in the future:

- Clinical data
- Administrative/claims data
- Patient-generated/reported data
- Non-traditional, health-related digital data sources

Clinical data. Patient-level data pulled from Electronic Medical Records (EMR) and clinical registries that describe how patients are treated. They include lab values, diagnoses, notes and not structured data, and other information from healthcare

Are Real World Data the smart way of doing Health Analytics?

visits with physicians.

Administrative/claims data. Patient-level data collected for non-clinical purposes, primarily for billing by providers to insurers and other payors, which can include diagnoses, services provided, costs, and other data required for the reimbursement of healthcare services. In general, the term “administrative” refers to information collected by government departments and other organisations primarily for administrative (i.e., not research) purposes.

Patient-generated/reported data. Individual data describing the patient’s experience.

Non-traditional, health-related digital data sources. As digital becomes increasingly prevalent in our lives, new sources of patient-level health data are emerging. These span social media posts that have a rich trove of information, especially health-focused social media sites.

The richness of information contained in such data is undeniable. It can be said that RWD holds potential to allow us to better design and conduct clinical trials and studies in the health care setting to answer questions previously though infeasible. All these statements are true, provided we unlock such potential and properly understand what it may and should be used for, and which methods should be developed and then applied to RWD.

3 Smart use of Real World Data in Healthcare Research

To unlock the potential of RWD, we have to point out i) who would benefit from their use; ii) which barriers their use may encounter, and iii) which are the question they are able to answer at best.

Starting from the latter point, RWD should be used to monitor the real use of healthcare (in terms of treatments, drugs, practices, healthcare paths, ...) pursued by a given population of interest. Such monitoring is needed to verify the applicability of guidelines arising from traditional clinical trials in a real world setting. In other words, RWD enables the evaluation of clinical practices and, in doing so, they represent a concrete support to decision and policy making.

The reasons why their use is still not intensive in healthcare research are many. Among the most significant barriers to expanding use of real-world data is the consensus that randomized controlled trials (RCT) remain the gold standard for demonstrating the efficacy and safety of medical products and treatments. This consensus, shared by physicians, patients, payors and regulators, creates significant hurdles to using RWD and to set a RWE-based decision making, even though there is a growing recognition that RCT alone cannot provide sufficient data for informed healthcare decision making in some situations. Limitations can also be attributed to a lack of

common technologies used across institutions that collect the initial patient data, and then to the uneven quality of such data. Moreover, the accessibility of RWD, which still represent one of the main issues for using them as basis in healthcare research, differs consistently from region to region and from country to country. Last but not least, the lack of standardization of RWE analytics makes stakeholders doubtful about the use of RWE in healthcare research.

This last point represent a strong limitation, since leads patients and clinicians not to understand that they would be the first and main beneficiaries of the use of RWD in clinical practice and healthcare research. In fact, the use of RWD in regulations and decision making is something that, in the end, would improve the quality of healthcare offered to patients and the efficiency of the system the clinicians work in.

4 What's next?

Expanding the use of RWE requires a multi-stakeholder action aimed at:

Increasing understanding and communication of RWE value. RWE analytics delivers valuable information, which healthcare researchers and data scientists are responsible for getting into the hands of payors, healthcare providers, and regulators to improve their healthcare decision making.

Shaping an integrated, adaptive partner ecosystem. Collaborations among academy, reference administrations and companies are essential to ensure credibility and trust in analyses, as well as gain access to novel data sources. Partnerships with database owners will be required in the short term to use data in restricted access databases, especially large, government-funded databases of public health systems in Europe which typically restrict access to this data.

Building platforms at scale to manage and analyze data in a rapid, low-cost fashion. RWE would incredibly benefit of platforms incorporating standardized methodology, which can be applied across all studies, improving the robustness and credibility of outputs.

5 Conclusions

Modern healthcare is undergoing a huge shift in how key stakeholders are approaching and evaluating patient data. An indicator for this shift is the growing access to and use of RWD, such as de-identified data collected from registries, electronic health records, wearable devices, and administrative and healthcare claims databases. RWE could significantly improve healthcare decisions across the health

Are Real World Data the smart way of doing Health Analytics?

system and ultimately improve patient care, provided we unlock the potential of RWD.

To achieve this goal, data integration and appropriate statistical techniques are essential ingredient, since the challenge of getting reliable evidence by RWD passes through:

- achieving medical-level accuracy;
- integrating data from multiple sources;
- collaborating with numerous stakeholders holding the data;
- removing confounders;
- developing suitable statistical methods;

complying the regulatory requirements of the context the data are generated from.

Making progress on this will establish the kind of culture where RWE is needed for making a new paradigm of healthcare research feasible.

References

1. Armitage, P., Berry, G., Matthews, J.N.: Statistical Methods in Medical Research, 4th Edition, Wiley-Blackwell (2001) ISBN: 978-0-632-05257-8
2. Corrao G., Mancia G.: Generating Evidence From Computerized Healthcare Utilization Databases. *Hypertension*, **65**: 490–498 (2015).
3. Ieva, F., Gasperoni, F.: Discussion of the paper "Statistical challenges of administrative and transaction data" by David J. Hand. *Journal of the Royal Statistical Society - Series A* (2018) doi: 10.1111/rssa.1231
4. Hand, D.J.: Statistical Challenges of administrative and transaction data. *Journal of the Royal Statistical Society - Series A* (2018) doi: <https://doi.org/10.1111/rssa.12315>
5. Hess L.M., Raebel M.A., Conner D.A., Malone D.C., Measurement of adherence in pharmacy administrative databases: a proposal for standard definitions and preferred measures. *Annals of Pharmacotherapy*, **40**: 1280–88 (2006)
6. Karve S., Cleve M.A., Helm M. et al., Prospective Validation of Eight Different Adherence Measures for Use with Administrative Claims Data among Patients with Schizophrenia. *Value In Health* **12**(6) (2009).
7. Mazzali, C. et al., Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in Lombardy region, 2000 to 2012. *BMC Health Service Research*, **16** (234) (2016).
8. Velentgas, P., Dreyer, N.A., Nourjah, P., Smith, S.R., Torchia, M.M., eds. Developing a Protocol for Observational Comparative Effectiveness Research: A Users Guide. AHRQ Publication No. 12(13)-EHC099. Rockville, MD: Agency for Healthcare Research and Quality; January 2013. www.effectivehealthcare.ahrq.gov/Methods-OCER.cfm.

Internet use and leisure activities: are all young people equal?

Internet e tempo libero: i giovani sono uguali tra loro?

Giuseppe Lamberti, Jordi Lopez Sintas and Pilar Lopez Belbeze

Abstract We explored how internet use could preclude or increase young people's in-person leisure activities and also explored how the relationship between internet use and leisure activities could vary by considering different sources of heterogeneity. We first investigated the internet use-leisure activities relationship globally estimating each as a single construct, and we then decomposed the indicators of both constructs in groups according to their nature, taking into account the different components of internet use and leisure activities. Our results suggest that the relationship is positive and stable. However, differences were identified once we considered the nature of the activities.

Abstract *In questo articolo, abbiamo analizzato l'impatto dell'uso d'internet su diverse attività di svago dei giovani, studiando anche come questa relazione varia considerando differenze demografiche e socio-economiche. L'uso d'Internet e le attività di svago sono state stimate sia globalmente (considerando due singole componenti), sia esaminando la diversa tipologia delle attività. I nostri risultati suggeriscono che la relazione è stabile e positiva. Tuttavia, delle differenze sono state identificate quando si analizza la relazione tenendo in conto la diversa natura delle attività.*

Keywords: Leisure, Internet, Pathmox, PLS-SEM, Heterogeneity

¹ Giuseppe Lamberti, Post-doctoral position at the business department of the University Autònoma of Barcelona AUB; email: giuseppe.lamberti@uab.cat

Jordi Lopez Sintas, Full professor position at the business department of the University Autònoma of Barcelona AUB; email: jordi.lopez@uab.cat

Pilar Lopez Belbeze, Professor position at the business department of the University Autònoma of Barcelona AUB; email: pilar.lopez@uab.cat

1 Introduction

Computers and the internet have become an integral part of leisure time. Technological advances have also provided people with more free time as well as increasing the range of leisure available to people. According to Dong (2017), technological changes have affected leisure activities both directly and indirectly, in particular by changing the nature of leisure practices and facilitating access to activities. The internet in particular has represented one of the more important technological advances in terms of affecting leisure activities. Despite efforts to understand the relationship between internet use and leisure activities, cumulative research results suggest that the situation is complex and controversial (Zach and Lissitsa, 2016). Evidence has indicated that interaction through the internet intensifies people's activities (Hwang, Cheong and Feeley, 2009). Rojas and Puig-i-Abril (2009) have reported that the internet provides almost unlimited opportunities, as an essential resource of information on education, business, shopping, travelling and healthcare. However, the internet may reduce in-person leisure activities as computers are mostly used in solitude and displace potentially more interactive social activities, as reported by Sagioglou and Greitemeyer (2014). Since research has been limited to a few indicators concerning a particular set of in-person leisure and internet activities, it is not possible to identify a fine-grained relationship between the two domains of internet use and leisure activities. Applying two different approaches, we explored the relationship between a comprehensive set of internet uses and young people's in-person leisure activities and also how this relationship varied by considering different sources of heterogeneity. Firstly, we investigated the relationship using a global approach, analysing internet use and leisure activities as two single constructs. Secondly, we decomposed the indicators of both constructs in groups according to their nature, taking into account the different components of internet use and leisure activities, and then fitted the effects between the different internet use scales to the identified leisure activity factors. This grouped approach enabled an understanding of interactions between different groups of activities when heterogeneity was considered. We used partial least squares structural equation modeling (PLS-SEM) (Hair et al., 2012) to study the relationship between the two sets of variables and Pathmox analysis (Lamberti, Aluja and Sanchez 2016, 2017) to explore the effect of heterogeneity.

2 Research design

We investigated the link between young people's conventional leisure practices and their uses of the internet using a database obtained from the 2017 Catalan Institute for Youth survey of a sample 3,152 young Catalans. Respondents were asked with what frequency they engaged in activities, reflecting their responses on a four-point Likert scale. Considered were 10 popular internet activities that have featured in recent scientific and market research (van Deursen and van Dijk, 2014) and 15 leisure activities (Jopp and Hertzog, 2007) (Table 1). As sources of heterogeneity we examined four categorical variables: age, gender, work status and income. PLS-SEM, which assumes the latent variables to be modelled as composites, directly allows latent variable proxies to be obtained as a linear combination of the manifest variables. Pathmox enables different segments of people to be identified that hold a different set of relationships among constructs. The method relies on binary segmentation principles to produce a tree with different models in each obtained node. The algorithm starts by fitting a global model to all the data to define the root of the tree and then, in a procedure that is repeated iteratively, identifies the models with the most significant differences in each child node. The available data is thus recursively partitioned to identify iterations whose segmentation variable yields the most significant difference.

Table 1: Internet use and leisure activities scales

Constructs/items		Constructs/items	
	<i>Internet use</i>		<i>Leisure activity</i>
Digital entertainment	Downloading movies	Culture/leisure	Reading
	Downloading software		Politics
Watching videos	Theatre		
Digital social interaction	Chatting		Concerts
Using social network sites	Museums		
Digital information and news	Reading online press	Social interaction	Clubbing
Listening/watching online radio/TV	Going out at night		
Information seeking	Meeting friends		
Commercial transactions	E-mailing	Outdoor leisure	Meeting friends
Shopping or ordering goods	Sports		
	Excursions		
	City walking		
	Travelling		
		Relaxation/study	Gaming
			Studying

3 Results

We first validated the measurement models. As both constructs were formative, we checked the weights and their significance using a bootstrap procedure with K=500 replicates and we verified multicollinearity by calculating the variance inflator factor. We then analysed the results for the structural models for the global approach and the grouped approach. The global approach results (Figure 1A) demonstrated that the two constructs were related and that internet use affected

leisure activities. The path coefficient of 0.488 indicated strong positive effects and this value was significant according to the bootstrap intervals. However, the R^2 value of 0.238 is considered a moderately small value for this kind of study. The grouped approach results (Figure 1B) identified how different categories of activities were related. Thus, the most important drivers for cultural leisure was digital information and news ($\beta=0.283$) and commercial transactions ($\beta=0.139$). Relationship maintenance was associated with social interaction ($\beta=0.312$) followed by digital entertainment ($\beta=0.156$). Regarding outdoor leisure, the most important driver was digital information and news ($\beta=0.119$). As for relaxation and study activities, the detected drivers were digital entertainment and commercial transactions ($\beta=0.234$ and $\beta=0.121$, respectively).

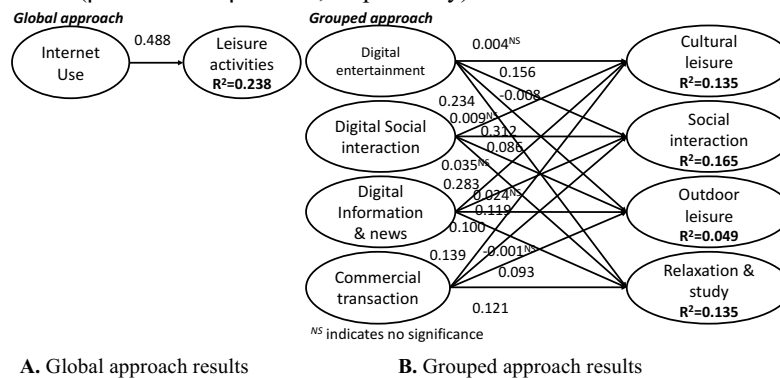


Figure 1: PLS-SEM structural model results.

3.1 Pathmox analysis

The Pathmox analysis was carried out using the available categorical variables as input variables in the segmentation procedure. Pathmox did not identify any significant partition for the global approach, whereas for the grouped approach, it identified three different groups of young people according to the *work status* and *gender* categorical variables, namely, students, male workers, and female workers. The algorithm identified *work status* as the variable with the greatest

Young people internet use and leisure activities

discriminating power ($p < 0.001$) and differentiated between students and others (inactive, employed and unemployed). This latter group was split according to *gender* ($p < 0.001$). The PLS-SEM models associated with the three groups are presented in Table 2.

Table 2: Local model results

<i>Path coefficients</i>	<i>Students</i>	<i>Male workers</i>	<i>Female workers</i>
Digital entertainment -> cultural leisure	-0.053 ^{NS}	0.059 ^{NS}	0.071
Digital social interaction -> cultural leisure	0.040 ^{NS}	-0.008 ^{NS}	-0.088
Digital information and news -> cultural leisure	0.297	0.314	0.279
Commercial transactions -> cultural leisure	0.092	0.088	0.205
Digital entertainment -> social interaction	0.031 ^{NS}	0.167	0.119
Digital social interaction -> social interaction	0.338	0.289	0.251
Digital information and news -> social interaction	0.099	-0.028 ^{NS}	0.048 ^{NS}
Commercial transactions -> social interaction	0.014 ^{NS}	-0.003 ^{NS}	0.095
Digital entertainment -> outdoor leisure	-0.036 ^{NS}	-0.054 ^{NS}	0.060 ^{NS}
Digital social interaction -> outdoor leisure	0.100	0.057 ^{NS}	0.012 ^{NS}
Digital information and news -> outdoor leisure	0.113	0.106	0.176
Commercial transactions -> outdoor leisure	0.133	0.091 ^{NS}	0.144
Digital entertainment -> relaxation/study activities	0.057 ^{NS}	0.274	0.118
Digital social interaction -> relaxation/study activities	0.032 ^{NS}	-0.008 ^{NS}	-0.011 ^{NS}
Digital information and news -> relaxation/study activities	0.180	0.117	0.135
Commercial transactions -> relaxation/study activities	0.163	0.110	0.201

^{NS} non-significant

4 Conclusion

Our analysis of the relationship between internet uses and young people in-person leisure activities makes two contributions. The PLS-SEM methodology not only allowed us to test the causal relationship between the two constructs but also to investigate which internet uses contribute to increasing young people's leisure activities. The Pathmox analysis of the effect of heterogeneity indicated that the relationship is not homogeneous but varies according to specific youth segments. We used two different approaches to analysis of the data, with the global approach indicating that internet use clearly affects leisure activities, and the grouped approach identifying how the different activities were related to each other. In particular, we identified that: (1) digital entertainment was the most relevant internet use for relaxation and studies, (2) digital social interaction was crucial to increasing the social interaction of young people, (3) digital information and news was essential to incrementing cultural activities, and (4) commercial transactions were relevant to both cultural leisure and relaxation and studies. When we tested the global model, we detected no significant differences according to the different segmentation variables. However, when we tested the grouped approach, we found three different groups demarcated according to the two categorical variables of work status and gender, with Pathmox defining three distinct segments. These findings reveal that the relationship between internet use and leisure activities did not vary according to the observed heterogeneity. However, when we considered how the different digital activities affected leisure activities, we detected

significant differences according to the three different groups (students, male workers, and female workers), meaning that heterogeneity mediates the effect of digital practices on leisure activities. In particular, we detected that for students, digital social interaction—such as chatting and using social media—were more relevant in defining social interactions (e.g., going out at night, meeting friends) and outdoor leisure activities—such as sports, excursion and city walking—than for the other two groups. For female workers, commercial transactions such as e-mailing, shopping and ordering goods had a greater effect on cultural leisure activities such as going to the theatre, concerts and museums. Finally, for male workers, digital entertainment in the form of movies, videos and software was significantly more important for male workers in defining social interactions and relaxation and study activities.

References

1. Dong, J.: Chinese elite migrants and formation of new communities in a changing society: An online-offline ethnography. *Ethnography*, 18(2), 221–239 (2017) <https://doi.org/10.1177/1466138116674225>
2. Hair, J.F., Sarstedt, M., Ringle, C.M. et al. *J. of the Acad. Mark. Sci.* (2012) 40 414. <https://doi.org/10.1007/s11747-011-0261-6>
3. Hwang, J. M., Cheong, P. H., & Feeley, T. H.: Being young and feeling blue in Taiwan: examining adolescent depressive mood and online and offline activities. *New Media & Society*, 11(7), 1101–1121 (2009) <https://doi.org/10.1177/1461444809341699>
4. Jopp, D., Hertzog, C.: Activities, self-referent beliefs and cognitive performance: Evidence for direct and mediated effects. *Psychology and Aging*, 22, 811–825 (2007) <http://dx.doi.org/10.1037/0882-7974.22.4.811>
5. Lamberti G., Banet Aluja T., and Sanchez G.: The pathmox approach for PLS path modeling: discovering which constructs differentiate segments, *Appl Stochastic Models Bus Ind.*, 33, 674-689 (2017) <https://doi.org/10.1002/asmb.2270>
6. Lamberti, G., Aluja, T. B., and Sanchez, G.: The pathmox approach for PLS path modeling segmentation, *Appl Stochastic Models Bus Ind.*, 32, 453-468 (2016) doi: 10.1002/asmb.2168
7. Rojas, H., & Puig-i-Abril, E.: Mobilizers mobilized: Information, expression, mobilization and participation in the digital age. *J Comput Mediat Commun.*, 14, 902-927 (2009) <https://doi.org/10.1111/j.1083-6101.2009.01475.x>
8. Sagioglou, C., & Greitemeyer, T.: Facebook's emotional consequences: why Facebook causes a decrease in mood and why people still use it. *Comput Human Behav.*, 35, 359-363 (2014) <https://doi.org/10.1016/j.chb.2014.03.003>
9. van Deursen, Alexander J.A.M. & Van Dijk, Jan A.G.M.: The digital divide shifts to differences in usage. *New Media & Society*. 16. 507-526 (2014) 10.1177/1461444813487959.
10. Zach, S., & Lissitsa, S.: Internet Use and Leisure Time Physical Activity of Adults - A Nationwide Survey. *Comput Human Behav.* 60 (2016) 10.1016/j.chb.2016.02.077.

On a Family of Transformed Stochastic Orders

Su una famiglia di ordinamenti stocastici trasformati

Tommaso Lando and Lucio Bertoli-Barsotti

Abstract We present a method to generalize basic stochastic orders through the transformation (referred to as probability distortion) of the cumulative probability distributions. In particular, we analyse the case of a power distortion function. The aim of this study is to obtain dominance relations that represent the preferences criteria in a more flexible way. The first and the second-degree stochastic dominance may be obtained as a limiting case and a special case, respectively.

Abstract *Presentiamo un metodo per generalizzare gli ordinamenti stocastici fondamentali tramite una trasformazione (la cosiddetta distorsione) da applicare alle funzioni di ripartizione. In particolare, viene studiato il caso in cui la funzione di distorsione è la potenza. L'obiettivo di questo studio è ottenere relazioni di dominanza che rappresentino in modo più flessibile i criteri di preferenza. Il primo e il secondo ordine di dominanza stocastica possono essere ottenuti rispettivamente come caso limite e come caso speciale.*

Key words: stochastic dominance, probability distortion

1 Introduction

Stochastic orders are primary tools for comparisons of random variables (RVs) with respect to some kind of preference relation (Shaked and Shanthikumar, 2007). Basic stochastic orders represent basic preferences such as “more” versus “less” and “less risky” versus “riskier”. Such concepts may be formalized in terms of inequalities between cumulative probability distributions (CDFs) or some integral transformation

¹ Tommaso Lando, University of Bergamo and VSB-Technical University of Ostrava; email: tommaso.lando@unibg.it

² Lucio Bertoli-Barsotti, University of Bergamo; email: lucio.bertoli-barsotti@unibg.it

of the CDFs, yielding the definitions of the first and second-degree stochastic dominance (FSD, SSD, respectively). Denote by F_X the CDF of an RV X .

Definition 1. We say that X FSD dominates Y and write $X \geq_1 Y$ iff

$$F_X(t) \leq F_Y(t), \forall t \in \mathbb{R}$$

Definition 2. We say that X SSD dominates Y and write $X \geq_2 Y$ iff

$$\int_{-\infty}^u F_X(t) dt \leq \int_{-\infty}^u F_Y(t) dt, \forall u \in \mathbb{R}.$$

Basically, FSD quantifies the concept of one RVs being “larger” than the other one, in some stochastic sense. Although all decision makers may agree with such idea, conditions for Def. 1 are generally too strong, so that it is rare to find distributions that can be ranked by FSD. In this regard, SSD makes it possible to increase the number of ranked distributions by introducing the concept of risk-aversion. Stated otherwise, all decision makers who prefer “more” to “less” and prefer “less risk” to “more risk” may conform to SSD.

Although FSD and SSD have been studied intensively in the literature, some authors recently focused back on this topic, stressing the fact that FSD is often unable to rank distributions, whereas SSD is generally too restrictive for those decision makers who are mostly risk-averse but may also have some slight attitude towards risk (Muller et al. 2016). This can be shown by the following simple example.

Example 1.

Let $X = c$, where $0 \leq c \leq 1$ and Y be a uniform RV defined on the support $[0,1]$. Clearly, X is less “risky” than Y , but the two CDFs always cross so that for $c = 1$ $X \geq_1 Y$ whilst $X \not\geq_1 Y$ for $c < 1$. If, $c \geq 0.5$ then $X \geq_2 Y$, whilst $c < 0.5$ then $X \not\geq_2 Y$. Nevertheless, we can argue that $c = 0.99$ should represent “almost all” decision makers, and this case is not equivalent to $c = 0.5$, although in both cases are represented by SSD.

Many different solutions have been proposed in the literature to solve the issue of finding intermediate dominance relations (Fishburn 1976, 1980, Leshno and Levy 2002, Tzeng et al. 2013, Tsetlin et al. 2015, Muller et al. 2016). We present a new approach based on a comparison of RVs whose distribution is transformed by a distortion function, that is, an increasing function $H: [0,1] \rightarrow [0,1]$ such that $H(0) = 0$ and $H(1) = 1$.

2 Transformed Stochastic Orders

Levi and Wiener (1998) studied dominance relations between transformed distributions, and investigated the classes of distortions that preserve the orderings.

On a Family of transformed Stochastic Orders

Following their approach, we compare RVs whose CDFs are transformed through a common distortion function H , and consequently define a family of stochastic orders parametrized by H . In particular we focus on a parametric class of distortion functions, i.e. the power family $H_k(t) = t^k$. We denote the corresponding dominance relation as *power-distorted stochastic dominance (PDS)*. Let X^k, Y^k be the RVs such that $F_{X^k} = (F_X)^k, F_{Y^k} = (F_Y)^k$.

Definition 3. We say that X PDS dominates Y with respect to H and write $X \geq_{1+1/k} Y$ iff $X^k \geq_2 Y^k$

We were able to prove some basic properties of PDS, summarized as follows.

- 1) $X \geq_{1+1/k} Y$ implies $X \geq_{1+1/m} Y$ iff $k \geq m$
- 2) $X \geq_{1+1/k} Y$ with $k = 1$ is equivalent to $X \geq_2 Y$
- 3) $X \geq_{1+1/k} Y$ is equivalent to $X \geq_1 Y$ when $k \rightarrow \infty$

Thus, PDS establishes a continuum of dominance relations, which generalizes FSD and SSD. PDS can be equivalently stated in terms of the quantile function. We recall that the quantile function of an RV X is given by

$$Q_X(p) = \inf\{z: F_X(z) \geq p\}$$

It can be seen that

$$X \geq_{1+1/k} Y \text{ iff } \int_0^u Q_X(t) dH(t) \geq \int_0^u Q_Y(t) dH(t) \text{ for all } u \in [0,1]$$

PDS is particularly simple to verify for discrete RVs. Let X be a discrete RV which takes values x_1, \dots, x_n . Thus the quantile of X is $Q_X(p) = x_i$ for $p \in I_i, i = 1, \dots, n$, where $I_i = [F_{i-1}, F_i) = [F_X(x_{i-1}), F_X(x_i))$ and $F_0 = Q_X(0)$, which can also be expressed as $Q_X(p) = \sum_{i=1}^n x_i 1_{I_i}(p)$, and $Q_X(1) = x_n$ (1_A is the indicator function of the set A). Then we obtain

$$\begin{aligned} \int_0^u Q_X(p) dp^k &= \int_0^u \sum_{i=1}^n x_i 1_{I_i}(p) k p^{k-1} dp = \\ &= \sum_{i=1}^{j-1} x_i ((F_i)^k - (F_{i-1})^k) + x_j (u^k - (F_{j-1})^k), \text{ for } u \in I_j. \end{aligned}$$

Example 2.

Let, for instance, X take the values 1,2,3 (with equal probabilities) and Y take the values 0,1,4 (with equal probabilities). It is easy to verify that $X \geq_2 Y$ although $X \not\geq_1 Y$ as the two CDFs cross. We can easily compute

$$\int_0^u Q_X(p) - Q_Y(p) dp^k = \begin{cases} t^k & 0 < t \leq \frac{2}{3} \\ 2^{1+k} 3^{-k} - t^k & \frac{2}{3} < t \leq 1 \end{cases}$$

and find $\int_0^u Q_X(p) - Q_Y(p) dp^k \geq 0, u \in [0,1]$, for $k \leq -\frac{\ln 2}{\ln 2 - \ln 3} = 1.7$. Thus $X \geq_{1.58} Y$.

Now, we turn back to Example 1 and apply PDSD.

Example 1. (Reprise)

We argue that for $c > 0.5$ the dominance relation between X and Y is stronger than SSD, whereas for $c < 0.5$ the dominance relation between X and Y is weaker.

$$\int_0^p Q_X(t) dt^k = c \int_0^p dt^k = c \cdot p^k$$

$$\int_0^p Q_Y(t) dt^k = kp^{k+1}/(k+1)$$

$\frac{\int_0^p Q_X(t) dt^k}{\int_0^p Q_Y(t) dt^k} = \frac{c(1+k)}{kp} \geq 1 \forall p \in [0,1]$ for $1 + \frac{1}{k} \geq \frac{1}{c}$, hence $X \geq_{1/c} Y$, i.e., the order of the PDSD is the reciprocal of c . This result confirms our conjecture. For instance, if $c = 0.6$, $X \geq_{1.6} Y$, if $c = 0.8$, $X \geq_{1.25} Y$, if $c = 0.99$, $X \geq_{1.01} Y$. Clearly, if $c = 1$, then $X \geq_1 Y$.

Funding

This research was supported by the Czech Science Foundation (GACR) under project 17-23411Y (to T.L.)

References

1. Fishburn, P. C. (1976). Continua of stochastic dominance relations for bounded probability distributions. *Journal of Mathematical Economics*, 3(3), 295-311.
2. Fishburn, P. C. (1980). Continua of stochastic dominance relations for unbounded probability distributions. *Journal of Mathematical Economics*, 7(3), 271-285.
3. Levy, H., & Wiener, Z. (1998). Stochastic dominance and prospect dominance with subjective weighting functions. *Journal of Risk and Uncertainty*, 16(2), 147-163.
4. Leshno, M., & Levy, H. (2002). Preferred by “all” and preferred by “most” decision makers: Almost stochastic dominance. *Management Science*, 48(8), 1074-1085.
5. Müller, A., Scarsini, M., Tsetlin, I., & Winkler, R. L. (2016). Between first- and second-order stochastic dominance. *Management Science*, 63(9), 2933-2947.
6. Shaked, M., and Shanthikumar, J. G. (2007). *Stochastic orders*. Springer Science and Business Media.
7. Tsetlin, I., Winkler, R. L., Huang, R. J., & Tzeng, L. Y. (2015). Generalized almost stochastic dominance. *Operations Research*, 63(2), 363-377.
8. Tzeng, L. Y., Huang, R. J., & Shih, P. T. (2013). Revisiting almost second-degree stochastic dominance. *Management Science*, 59(5), 1250-1254.

Bayesian stochastic search for Ising chain graph models

Ricerca stocastica Bayesiana per modelli grafici a catena Ising

Andrea Lazzerini · Monia Lupparelli · Francesco C. Stingo

Abstract A class of Ising chain graph models is illustrated to explore the effect of an external risk factor on a set of binary outcomes and on their joint dependence structure modelled via undirected graphs, where we are mainly interested in the risk factor effect on pairwise associations between outcomes rather on single outcomes. Under the LWF Markov property, the joint probability mass associated to a chain graph corresponds to a log-linear model with suitable zero constraints in correspondence of missing edges. We devise a Bayesian Ising model based on conjugate priors for log-linear parameters that aims at the selection of the best graph that fits the data. A computational strategy is implemented that uses Laplace approximations and a Metropolis-Hastings algorithm that allows us to perform model selection.

Abstract Si illustra una classe di modelli grafici Ising per esplorare l'effetto di un fattore di rischio esterno su un insieme di risposte binarie e sulla loro struttura di dipendenza modellata con grafi non orientati, dove l'interesse principale riguarda l'effetto del fattore di rischio sulle associazioni a coppia delle risposte. Date le proprietà Markoviane LWF, la funzione di probabilità congiunta di un grafo a catena corrisponde a un modello log-lineare con vincoli a zero per gli archi mancanti. Con un approccio Bayesiano, utilizziamo un modello Ising basato su distribuzioni a priori coniugate per i parametri che mirano alla selezione del grafo che meglio si adatta ai dati. Si implementa una strategia che usa l'approssimazione di Laplace e un algoritmo di Metropolis-Hastings che permette di fare selezione del modello.

Key words: log-linear models, Diaconis and Ylvisaker prior, Bayesian model selection

Andrea Lazzerini

University of Bologna, Via Zamboni, 33 - 40126 Bologna, e-mail: andrea.lazzerini2@unibo.it

Monia Lupparelli

University of Florence, Piazza di San Marco, 4 - 50121 Firenze, e-mail: monia.lupparelli@unifi.it

Francesco C. Stingo

University of Florence, Piazza di San Marco, 4 - 50121 Firenze, e-mail: francescoclaudio.stingo@unifi.it

1 Discrete chain graphs and LWF Markov Properties

A graphical Markov model is a statistical model defined over a graph whose vertices correspond to random variables. The missing edges of the graph are translated into conditional independence restrictions that the model imposes on the joint distribution of the variables.

We are interested in chain graphical models where the graph may have both directed and undirected edges under the constraint that there do not exist any semi-directed cycles. Let $G = (V, E)$ be a chain graph, where V is the vertex set and E is the edge set: the absence of semi-directed cycles implies that V can be partitioned into so-called chain components $\tau \in T(G)$ such that edges within a chain component are undirected whereas the edges between two chain components are directed and point in the same direction.

The rules that govern how a graph is translated into conditional independence restrictions are known as Markov properties. Four classes of Markov properties for chain graphs have been discussed in the literature [4]. We focus on chain graphs under the LWF Markov properties [6]. The resulting models have log-linear structure which yields that the models are smooth exponential families.

Let $Y_V = (Y_\tau) : \tau \in T(G)$, be the random vector corresponding to the chain graph $G = (V, E)$; under several assumptions [6] and if the probability distribution $P(Y_V)$ is strictly positive, then it holds the following factorization:

$$P(Y_V) = \prod_{\tau \in T(G)} P(Y_\tau | Y_{pa(\tau)}), \quad (1)$$

where we say that τ_a is a parent of τ_b , e.g. $pa(\tau_b) = \tau_a$, if at least one element in τ_b is reached by an arrow from at least one element in τ_a , with $\tau_a, \tau_b \in T(G)$.

An example including all the independencies specified by the chain graph model under the LWF Markov property is given in Fig 1.

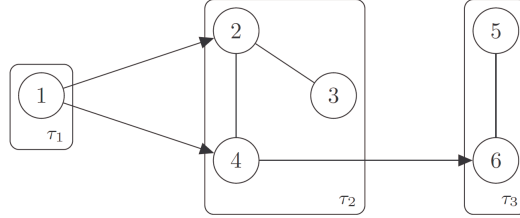


Fig. 1 An example of chain graph $G = (V, E)$ where $V = \{1, 2, 3, 4, 5, 6\}$, $E = \{(1, 2), (1, 4), (2, 3), (2, 4), (4, 6), (5, 6)\}$, $T(G) = \{\tau_1(1), \tau_2(2, 3, 4), \tau_3(5, 6)\}$. Under the LWF Markov properties it holds that $P(Y_V) = P(Y_1) P(Y_2, Y_3, Y_4 | Y_1) P(Y_5, Y_6 | Y_2, Y_3, Y_4)$ and $Y_3 \perp\!\!\!\perp Y_4 \mid Y_1, Y_2$

2 The Ising model and the prior distribution

2.1 The model

Let Y_V be the random vector corresponding to V , a set of p binary attributes, and X be the random variable corresponding to a binary factor external to V , taking value $x = \{0, 1\}$. Assume that we have observed $N^{(x)}$ realizations of $Y_V | X = x \sim \text{Multinomial}(N^{(x)}, \pi^{(x)})$ with $x = \{0, 1\}$. We denote with $n_i^{(x)}$ and $\pi_i^{(x)}$ the joint count and the joint probability corresponding to the i -th cell of the contingency table $I^{(x)} = \times (0, 1)^p$ for $Y_V | X = x$, with $i \in I^{(x)}$ and $x = \{0, 1\}$.

Probabilities are easy to interpret but have the drawback that submodels of interest typically involve nonlinear constraints on these parameters, as in the case of pairwise conditional independence relationships which can be specified by requiring certain factorizations of the cell probabilities. For instance, for $V = \{1, 2, 3\}$ it holds

$$Y_1 \perp\!\!\!\perp Y_2 \mid Y_3, X = x \iff \pi_{110}^{(x)} = \frac{\pi_{100}^{(x)} \pi_{010}^{(x)}}{\pi_{000}^{(x)}} \wedge \pi_{111}^{(x)} = \frac{\pi_{101}^{(x)} \pi_{011}^{(x)}}{\pi_{001}^{(x)}}, \quad (2)$$

where $\pi_{ijz}^{(x)} = P(Y_1 = i, Y_2 = j, Y_3 = z | X = x)$ with $x = \{0, 1\}$.

For this reason, it is useful to develop alternative parameterizations such that submodels of interest correspond to linear subspaces of the parameter space of the saturated model. In order to obtain an alternative reparameterization, that is the log-linear parameter λ , a fundamental role is played by the Zeta matrix (Z) and the Möbius matrix (M) [9], $M = Z^{-1}$, where

$$\lambda^{(x)} = M^T \log \pi^{(x)} \iff \pi^{(x)} = \exp(Z^T \lambda^{(x)}), \quad (3)$$

therefore, the log-linear expansion of $\pi^{(x)}$ can be used to express pairwise independencies on $Y_V | X = x$ as zero-constraints on $\lambda^{(x)}$ with $x = \{0, 1\}$; indeed (2) becomes

$$Y_1 \perp\!\!\!\perp Y_2 \mid Y_3, X = x \iff \lambda_{12}^{(x)} = 0 \wedge \lambda_{123}^{(x)} = 0, \quad (4)$$

with $x = \{0, 1\}$.

An issue about modelling categorical data is that as the number of the variables increases, the number of parameters can become so large to be intractable. A possible solution to this problem is to assume the Ising model, where all higher than 2-factor log-linear interactions vanish.

Then, let $y_j^{(x)}$ be the marginal count conditional on $X = x$ of the j -th element of $\Omega_V = \{a, b\}_{a, b \in V}$; if $Y_V | X = x \sim \text{Ising}(\lambda^{(x)})$,

$$\begin{aligned}
 P(Y_V|X=x) &\propto \prod_{i \in I^{(x)}} \pi_i^{(x) n_i^{(x)}} \\
 &= \exp \left\{ \sum_{j \in \Omega_V} y_j^{(x)} \lambda_j^{(x)} - N^{(x)} \log \left(1 + \sum_{i \in I^{(x)} \setminus i_\emptyset^{(x)}} \exp \sum_{j \in \Omega_i} \lambda_j^{(x)} \right) \right\} \quad (5)
 \end{aligned}$$

where $i_\emptyset^{(x)}$ is the baseline cell conditional on $X = x$ with $x = \{0, 1\}$. Note that if $|j| = 1$, $\lambda_j^{(x)}$ represents the effect of $X = x$ on the single variable, while if $|j| = 2$, $\lambda_j^{(x)}$ represents the association between the two variables for $X = x$ with $x = \{0, 1\}$. The comparison of $\lambda_j^{(x)}$ for $x = \{0, 1\}$ represents the effect of X on the association between the two variables.

In this framework, our general aim is to perform a Bayesian model selection of the chain graph $G = (\{V, X\}, E)$, with $T(G) = \{\tau_1(X), \tau_2(V)\}$ and $P(Y_V, X) = P(X) P(Y_V|X)$.

2.2 Prior distribution

In the literature several methods have been proposed aimed at the selection of the best log-linear model; concerning the frequentist approach, most of them fall into one of the following three classes of methods: multiple testing strategies [1], screening procedures [5] and regularization methods [8].

Regarding the Bayesian approach, the strategies for finding models with high posterior probability can be classified by the choice of the prior distribution of the parameters where we recall three possibilities: the conjugate Hyper-Dirichlet prior on π (conjugate only for decomposable graphs) [2], the multivariate Normal not conjugate prior on λ [3] and the Diaconis and Ylvisaker conjugate prior on λ [7].

To select the best Ising graph model we decide to follow a Bayesian approach using the Diaconis and Ylvisaker prior [7], which has the following form

$$\begin{aligned}
 f(\lambda^{(x)}|m^{(x)}) &= C(s^{(x)}, \alpha^{(x)})^{-1} \\
 &\times \exp \left\{ \sum_{j \in \Omega_V} \lambda_j^{(x)} s_j^{(x)} - \alpha^{(x)} \log \left(1 + \sum_{i \in I^{(x)} \setminus i_\emptyset^{(x)}} \exp \sum_{j \in \Omega_i} \lambda_j^{(x)} \right) \right\} \quad (6)
 \end{aligned}$$

where, $C(s^{(x)}, \alpha^{(x)})$ is the normalising constant and $s^{(x)} \in \mathbb{R}^{k^{(x)}}$, $\alpha^{(x)} \in \mathbb{R}$ are the hyperparameters with $k^{(x)}$ the number of parameters in the model $m^{(x)} \in M^{(x)}$ with $x = \{0, 1\}$.

3 Posterior inference

Let $m^{(x)}$ be a model selected from a set of competing models $M^{(x)}$ and $n^{(x)}$ be the data; we denote with $f(m^{(x)})$ and $f(m^{(x)}|n^{(x)})$ the prior and the posterior probability of $m^{(x)}$. Assuming that all the models are a priori equally likely, the posterior probability of $m^{(x)}$ is proportional to the marginal likelihood $f(n^{(x)}|m^{(x)})$, i.e.

$$f(m^{(x)}|n^{(x)}) \propto f(n^{(x)}|m^{(x)}) = \int_{\Theta_{m^{(x)}}} f(\lambda^{(x)}|m^{(x)}) f(n^{(x)}|m^{(x)}, \lambda^{(x)}) d\lambda^{(x)} \quad x = \{0, 1\}, \quad (7)$$

where $f(\lambda^{(x)}|m^{(x)})$ is the prior probability of the parameters $\lambda^{(x)} \in \Theta_{m^{(x)}}$ and $f(n^{(x)}|m^{(x)}, \lambda^{(x)})$ is the likelihood.

Using the Diaconis and Ylvisaker prior, the integral in (7) is analitically derived as

$$\frac{C(y^{(x)} + s^{(x)}, N^{(x)} + \alpha^{(x)})}{C(s^{(x)}, \alpha^{(x)})} \quad x = \{0, 1\}, \quad (8)$$

the ratio between the normalising constants of the posterior and prior distributions of the parameters [7]. We calculate both the normalizing constants through the Laplace approximation [10] such that

$$C = Ke(\lambda^{*(x)}) \frac{(2\pi)^{k^{(x)}/2}}{|A^{(x)}|^{1/2}} \quad x = \{0, 1\}, \quad (9)$$

where $Ke(\lambda^{*(x)})$ is the kernel of the function to be approximated and $A^{(x)}$ is the Hessian matrix ($k^{(x)} \times k^{(x)}$), both evaluated in a stationary point $\lambda^{*(x)}$.

Since the size of the set of possible models $M^{(x)}$ is too large to be explored entirely, we perform a stochastic model search. Starting from $m^{(x)(t-1)}$, the model accepted at time $(t-1)$, we propose a new model $m^{(x)(t)}$ by randomly selecting one log-linear parameter (except those of the main effects) and by adding or removing it from the model. We finally accept or reject $m^{(x)(t)}$ with a Metropolis-Hastings step.

4 Simulation study

To evaluate the method efficiency we perform a simulation study where our goal is to select the best chain graph model $G = (\{V, X\}, E)$, with $p = 10$. Therefore, for $X \in \{0, 1\}$ we generate $N^{(x)} = 2500$ observations from $Y_V|X = x \sim \text{Ising}(N^{(x)}, \lambda^{(x)})$ where $\lambda^{(x)} \sim N_k(0, I_k)$ with $x = \{0, 1\}$, k equal to the 20% of the parameters of the saturated model and I_k the identity matrix. We report the results in Fig. 2, concerning the boxplots of the percentage of interactions correctly identified for 10 different models, with 1000 MCMC steps for each.

We finally remark that our interest is mainly focused in modelling the effect of the risk factor of the dependence structure of the attributes rather than on single attributes. In terms of model search, this issue requires to explore a larger model space and therefore we need to better investigate suitable parameterizations and selection strategies which allow us to better perform a compatible procedure for model choice.

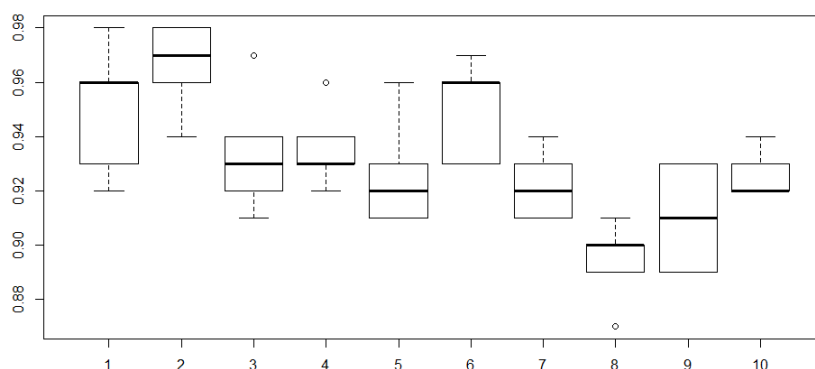


Fig. 2 Boxplots of the percentage of interactions correctly identified for 10 different models, with 1000 MCMC steps for each.

References

1. Aitkin, M.: A note on the selection of log-linear models. *Biometrics* **36**, 173–178 (1980)
2. Dawid, A.P., and Lauritzen, S.L.: Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Stat.* **21**, 1272–1317 (1993)
3. Dellaportas, P., Forster, J.J.: Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86**, 615–633 (1999)
4. Drton, M.: Discrete chain graph models. *Bernoulli* **17**, 736–753 (2009)
5. Edwards, D., Havr nek, T.: A Fast Procedure for Model Search in Multidimensional Contingency Tables. *Biometrika* **72**, 339–351 (1985)
6. Frydenberg, M.: The Chain Graph Markov Property. *Scand. J. Stat.* **17**, 333–53 (1990)
7. Massam, H., Liu J., Dobra, A.: A conjugate prior for discrete hierarchical log-linear models. *Ann. Stat.* **37**, 3431–3467 (2009)
8. Ravikumar, P. et al.: High-dimensional Ising model selection using l_1 -regularized logistic regression. *Ann. Stat.* **38**, 1287–1319 (2010)
9. Roverato, A.: *Graphical Models for Categorical Data (SemStat Elements)*. Cambridge: Cambridge University Press (2017)
10. Tierney, L., Kadane, J.: Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.* **81**, 82–86 (1986)

On the statistical design of parameters for variables sampling plans based on process capability index C_{pk}

Progettazione statistica dei parametri per il piano di campionamento per variabili basato sull'indice di capacità di processo C_{pk}

Antonio Lepore, Biagio Palumbo and Philippe Castagliola

Abstract This paper deals with the statistical design of the plan parameters (viz., required sample size and critical acceptance value) recently proposed in the literature for variables sampling plan based on process capability index C_{pk} . The use of incorrect plan parameters leads in fact to a considerably lower probability of acceptance than the one desired by the producer. This issue occurs in the case of sampling plan for both single and resubmitted lots. In order to illustrate practical consequences, a real example in the aerospace manufacturing process of a key characteristic is eventually presented.

Abstract Questo lavoro riguarda la progettazione statistica dei parametri (viz., ampiezza campionaria minima e valore critico di accettazione) per il piano di campionamento basato sull'indice di capacità di processo C_{pk} recentemente proposto in letteratura. Viene mostrato come l'uso di parametri non adeguatamente progettati determini una minore probabilità di accettazione del lotto considerato, rispetto a quella imposta dal produttore. Tale problema interviene sia nel caso di piani di campionamento con reimmissione che senza reimmissione. Per descriverne le conseguenze pratiche, viene infine presentato un caso reale su dati provenienti da un processo reale nell'industria aerospaziale.

Key words: Quality control; Process capability indices; Critical acceptance values; Process capability analysis; Lot resubmission.

Antonio Lepore and Biagio Palumbo
Department of Industrial Engineering, University of Naples Federico II, Naples, Italy
e-mail: antonio.lepore@unina.it; biagio.palumbo@unina.it

Philippe Castagliola
Université de Nantes & LS2N UMR CNRS 6004, Nantes, France
e-mail: philippe.castagliola@univ-nantes.fr

1 Introduction

Process monitoring and quality control play a role in today's aerospace manufacturing field. Control charts and process capability indices (PCIs) are essential to monitor key product dimensional characteristics and measure the process capability to meet the specification interval, respectively. A large literature devoted to PCIs has been produced during the last decades that covers both statistical issues and industrial practices (see e.g., [3]).

Rather than evaluating all the items of a large production lot, a specified sample is taken, inspected or tested, and a decision is made about accepting or rejecting the entire production lot. The sampling naturally involves risks of not adequately reflecting the quality conditions of the lot. In this setting, the Type I (resp. Type II) risk is the probability of incorrectly rejecting (resp. accepting) a lot that is acceptable (resp. unacceptable) and is also called producer's (resp. consumer's) risk or α -risk (resp. β -risk). A *variables sampling plan* is a statement regarding the required sample size for product inspection and the associated acceptance or rejection criteria for lot disposition based on a quality characteristic that is measured in a continuous scale. It provides both producers and consumers to reserve their own rights by compromising on a rule to judge a lot. A well-designed sampling plan can significantly reduce the difference between the required and the actual product quality of the lot.

The single acceptance sampling plan is based on a single sample and is the most used plan because of its simplicity. However, when a lot is rejected based on a single sample, the producer may dispute the first lot disposition and he is allowed to resample the same number of units under the provisions of the contract. Even though it makes sense to combine the results of the two samples, as is done in the traditional double sampling plan, the practice in these situations is to discard the first sample and to use the second. In general, by extending this strategy, [5, 7] allows for resampling –and thus, resubmission– (at most) m times.

In this background, acceptance sampling plans, which are popularized in [1], can be based on PCIs as practical tool for decision making about accepting or rejecting a production lot based on a specific sample drawn from it. However, noticeable resistance still remains in considering the variability of PCI point estimators that can lead to improper corrective actions in practice.

In last decades, PCIs have received increasing attention in the quality assurance and statistics fields. Based on PCIs, a production department can identify and improve a poor process so that the product quality can be enhanced and the requirements of the customers can be met. In a purchasing contract, a threshold value has usually to be specified for PCIs. If the PCI exceeds the given threshold, the process is judged as incapable. Otherwise, the process is said to be capable. The C_{pk} index [2] is the most widely used PCI in industrial practice as it is the simplest PCI able to account for both the magnitude of process variation and yield.

Specifically, if a characteristic X is normally distributed with mean μ and standard deviation σ and the target value T is assumed as the midpoint of the specification interval $[LSL, USL]$, the C_{pk} index is defined as

$$C_{pk} = \frac{\min(\mu - LSL; USL - \mu)}{3\sigma}. \quad (1)$$

If the process can be demonstrably stable (i.e., *in control*), the natural PCI estimators of C_{pk} is denoted by \hat{C}_{pk} and obtained by replacing μ and σ in Eq. (1) with $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and $\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$, respectively, and by assuming the data drawn from the quality characteristic X as independent.

Specifically, design of the required sample size and the critical acceptance value (i.e. plan parameters) for variables sampling plans based on C_{pk} that are proposed in several published papers [5, 8, 7] are found inappropriate. In particular, it is shown to lead to probability of acceptance of an acceptable lot that is lower than the one desired by the producer ($1 - \alpha$), as demonstrated by the results appeared in [4]. Hence, plan parameters tabulated in [7] should not be used. In order to illustrate practical consequences, a real example in the manufacturing process of a key characteristic of an aerospace engine component is eventually presented.

2 Distribution of \hat{C}_{pk} and dependencies with variables sampling plans

If the quality characteristic X follows a normal distribution, an exact form of the distribution of C_{pk} exists [6]

$$F_{\hat{C}_{pk}}(y; n, C_{pk}, \xi) = 1 - \int_0^{(3C_{pk} + |\xi|)\sqrt{n}} G\left(\frac{(n-1)[(3C_{pk} + |\xi|)\sqrt{n} - t]^2}{9ny^2}\right) \times [\phi(t + \xi\sqrt{n}) + \phi(t - \xi\sqrt{n})] dt; \quad y > 0 \quad (2)$$

being G the cumulative distribution function of a chi-square random variable with $n - 1$ degrees of freedom, ϕ the standard normal probability density function and $\xi = (\mu - T)/\sigma$.

The crux of the matter is that this function does not depend only on the actual value of C_{pk} and the sample size n , but also on the additional parameter ξ , which is not known at given C_{pk} . It is useful to note that the function in Eq. (2) is an even function of ξ and thus we shall indistinctly refer to ξ and $|\xi|$. Given $\alpha, \beta \in (0, 1)$ and $m \in \mathbb{N}$, the optimal plan parameters $n \in \mathbb{N}$ and $C_0 > 0$ are defined as the smallest n and corresponding C_0 such that

$$\sup_{\xi} 1 - \left[F_{\hat{C}_{pk}}(C_0; n, C_{LTPD}, \xi) \right]^m \leq \beta \quad (3)$$

and

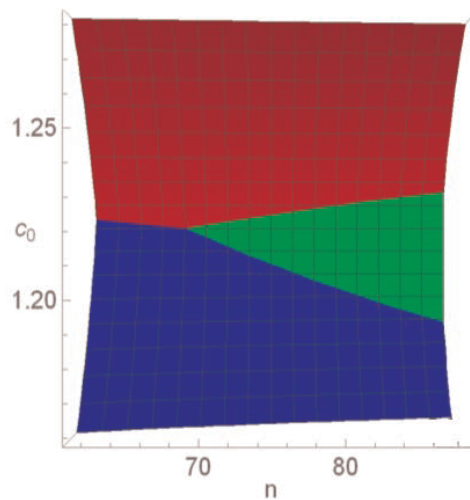
$$\sup_{\xi} \left[F_{\hat{C}_{pk}}(C_0; n, C_{AQL}, \xi) \right]^m \leq \alpha \quad (4)$$

In Eq. (3), C_{LTPD} is the *limiting quality level* (sometimes referred to also as *lot tolerance percent defective*) at which the consumer desires the probability of acceptance

being no more than β ; whereas, in Eq. (4), C_{AQL} is the *acceptable quality level* at which the producer desires a (high) probability of acceptance at least equal to $1 - \alpha$ ($C_{LTPD} < C_{AQL}$). In this regard, the recommendation found through numerical investigation in [7] to simultaneously solve Eq. (3) and Eq. (4) by setting $\xi = 1$ is theoretically disproved in [4]. In particular, it turns out that Eq. (4) has to be solved by setting $\xi = 0$. So stated, the plan parameters appeared in [7] for the resubmitted lots ($m > 1$) and in [5] for the variables single sampling plan ($m = 1$) are therefore, not conservative for the producer. In [4] the correct plan parameters are then recalculated and reported for practitioners' convenience and the error associated is shown to be increasing with respect to m . As an example with usual values of $C_{AQL} = 1.33$, $C_{LTPD} = 1.00$ and $\alpha = 0.05$, the true α -risk is shown to be larger than the double of the maximum one desired by the producer when $m = 10$.

In the light of this, in Figure 1 the solutions of the simultaneous equations (3) and (4) with $\xi = 1$ and $\xi = 0$, respectively, given $\alpha = \beta = 0.05$ are represented in a $n - C_0$ plane as the infinite triangular-shaped region in green. Accordingly, the same figure reports in red (resp. blue) the region where the producer's (resp. consumer's) risk is larger than α (resp. β). The real solution is the vertex of this region. However, the minimum sample size n has to be an integer value and thus usually has to move away from the vertex. From Figure 1 it is then clear that —even having fixed the levels α and β — by rounding n to the smallest integer greater than the real solution (depicted on the vertex of the green region of Figure 1) C_0 is not unique and allowed to be arbitrarily chosen in the interval along C_0 -axis having extremes at the boundary of the red and blue regions. As far as we know, this aspect has never been pointed out in the literature, even if, in some critical cases, it impacts considerably on the true consumer's and producer's risks.

Fig. 1 Graphical representation of solutions (green region) of the simultaneous equations (3) and (4)



3 Real example and conclusion

In order to illustrate another yet subtle drawback when using inappropriate plan parameters, a real dataset from a manufacturing process of the key characteristic `diameter` of an aerospace engine component is analysed. The characteristic under study has been preliminarily demonstrated to be under statistical control through control charts for variables with respect to a Phase I historical sample. In this case, industrial practitioners use the variables single ($m = 1$) sampling plan with $C_{AQL} = 1.33$, $C_{LTPD} = 1.00$, and $\alpha = \beta = 0.1$. In view of the (minimum) required sample size $n = 49$ suggested in [5, 7], the actual sample size is set equal to $n^* = 54$. Accordingly, by following [5, 7], practitioners are misleadingly able to find values for C_0 . However, we observe the correct sample size should exceed $n_C = 57$ as proposed in [4] and thus, in fact, no solution is available for (n, C_0) that satisfies both producer's and consumer's specifications.

By following the illustration reported in Figure 1, n^* falls on the right of the vertex of the green region. That is, whatever the critical acceptance value C_0 , α -risk $\in [0.115, 0.149]$, in spite of the maximum one ($\alpha = 0.1$) desired by the producer. In other words, the actual sample size ($n^* = 54$) designed for a lot disposition through the variables single sampling plan used in the production line of the `diameter` is not adequate, even if higher than the minimum required reported in [5, 7]. It must necessarily be increased at least to $n_C = 57$; otherwise, the actual producer's α -risk results inevitably larger than the maximum desired by the producer.

In a nutshell, this real example underlines that when using a variables sampling plan based on C_{pk} , if the sample size is smaller than the one reported in [4], any critical acceptance value does not achieve the desired producer's risk α . Therefore, in designing parameters for variables sampling plans based on C_{pk} , practitioners should rely only on the tables reported in [4] to set the minimum sample size and the corresponding critical acceptance value C_0 . As future directions, even if only by numerical investigation, we found similar issues also in variables sampling plan parameters proposed in [8] and based on the index C_{pmk} , even if, practically speaking, they leads to negligible inaccuracy at commonly used values of α , β , C_{AQL} and C_{LTPD} .

References

1. Dodge, H.F., Romig, H.: A method of sampling inspection. The Bell System Technical Journal **8**(4), 613–631 (1929)
2. Kane, V.E.: Process capability indices. Journal of Quality Technology **18**(1), 41–52 (1986). DOI 10.1080/00224065.1986.11978984. URL <https://doi.org/10.1080/00224065.1986.11978984>
3. Lepore, A., Palumbo, B.: New insights into the decisional use of process capability indices via hypothesis testing. Quality and Reliability Engineering International **31**(8), 1725–1741 (2015)
4. Lepore, A., Palumbo, B., Castagliola, P.: A note on decision making method for product acceptance based on process capability indices cpk and $cpmk$. European Journal of Operational Research **267**(1), 393–398 (2018)

5. Pearn, W.L., Wu, C.W.: An effective decision making method for product acceptance. *Omega* **35**(1), 12–21 (2007)
6. Vännman, K.: Distribution and moments in simplified form for a general class of capability indices. *Communications in statistics-theory and methods* **26**(1), 159–179 (1997)
7. Wu, C.W., Aslam, M., Jun, C.H.: Variables sampling inspection scheme for resubmitted lots based on the process capability index cpk. *European Journal of Operational Research* **217**(3), 560–566 (2012)
8. Wu, C.W., Pearn, W.L.: A variables sampling plan based on cpmk for product acceptance determination. *European Journal of Operational Research* **184**(2), 549–560 (2008)

Nowcasting foreign tourist arrivals using Google Trends: an application to the city of Florence, Italy

Nowcasting degli arrivi turistici stranieri usando Google Trends: un'applicazione nella città di Firenze, Italia

Alessandro Magrini

Abstract Google Trends (GT) data have been shown to improve nowcasting of tourism volume by several empirical researches in recent years. Unfortunately, a well known problem of GT is that the same query may provide different data when performed in different occasions. In this paper, we focus on foreign tourist arrivals in the city of Florence, Italy, and explore the extent to which GT data can be used to improve forecast across several replications of the same query, performed two times per day for three weeks. We found that only a part of the replications improves forecast, but the percentage of improvements is significantly increased if they are averaged for at least three consecutive days of download and filtered using 2x12 moving average.

Abstract *Varie ricerche empiriche negli ultimi anni hanno mostrato che i dati di Google Trends (GT) sono in grado di migliorare il nowcasting del volume del turismo. Purtroppo, un noto problema di GT è che la stessa query può fornire dati diversi se effettuata in diverse occasioni. In questo articolo ci si concentra sugli arrivi di turisti stranieri nella città di Firenze, in Italia, per esplorare in che misura i dati di GT possano migliorare le previsioni su più repliche della stessa query, eseguita due volte al giorno per tre settimane. Si è trovato che solo una parte delle repliche migliora le previsioni, ma la percentuale di miglioramenti aumenta significativamente se esse vengono mediate per 3 giorni consecutivi di download e filtrate con media mobile 2x12.*

Key words: Google Search index, time-series filtering, seasonal ARIMAX, time-series forecasting, tourism demand.

Alessandro Magrini

Dep. Statistics, Computer Science, Applications – University of Florence, Italy
e-mail: alessandro.magrini@unifi.it

1 Introduction

‘Nowcasting’ is a quite recent term introduced to enforce the concept of forecasting when the target is the present or the near future, a task strongly motivated by the increasing availability of search volume indices like Google Trends (GT). In particular, GT data have been claimed to improve nowcasting of tourism volume by several empirical researches in recent years (see [4] for a review).

Unfortunately, the sampling procedure of GT is not clear and entails several practical issues, in particular the same identical query may provide different data when performed in different occasions. The main technical issues of using GT data for nowcasting were highlighted by [1]. Firstly, it is pointed out that GT data are not a probabilistic sample, but a self-selected sample created by the internet users, thus they contain systematic bias that cannot be resolved through seer sample size. Secondly, it is highlighted that GT data can be seen as time series drawn from a repeated survey with unknown design, since the same internet users will supposedly make their research repeated in the short term, thus there is an overlap with adjacent observations inducing serial correlation. Thirdly, it is observed that GT data contains seasonality, which, combined with serial correlation, may appear more variable than it is, thus entailing bias.

In this paper, we explore the effectiveness of two techniques to deal with the sampling procedure of GT: averaging several replications of the same query and filtering each time series using 2x12 moving average. At this purpose, we consider the query ‘Visit Florence’ relative to the period 2008–2017, with data downloaded two times per day for three weeks, and use a seasonal Auto-Regressive Integrated Moving Average (ARIMA) model (see, for example, [3] Chapter 5) fitted to official data as benchmark. ARIMA models are widely used in the literature to represent time series of tourism phenomena, motivated by their effective performance in the tourism forecasting competition by [2].

This paper is structured as follows. In Section 2, a description of the data and the methodology of the research is provided. In Section 3, results are presented. Section 4 includes the discussion of the contribution.

2 Materials and methods

Official monthly data on foreign tourist arrivals (all type of accommodation) for the metropolitan area of Florence in the period 2008–2016 were downloaded from the ISTAT database (<http://dati.istat.it/>), and shown in Figure 1.

Search volume data were downloaded from the GT website (<https://trends.google.it/trends/>) two times per day for the first three weeks of February 2019, with query: ‘Visit Florence’, time period: 2008–2017 (monthly), language: English, thematic category: ‘travel’, geographic area: ‘whole world’. This led to a total of 42 different time series, shown in Figure 2. The cross-sectional coefficient

of variation (Figure 3) highlights decreasing level of noise in GT data until 2013, which appears stable in the successive years.

We fitted a seasonal Auto-Regressive Integrated Moving Average (ARIMA) model to data on foreign tourism arrivals (logarithmic scale), which was considered as the benchmark model to be compared to several augmented models. Each augmented model was obtained by adding to the benchmark a GT time series at lag 0 (same month) and lag 1 (previous month) as covariates. We considered 42 augmented models including each replication of GT data with no transformation, and 42 augmented models including each replication of GT data filtered using 2x12 moving average. Also, we considered 7 augmented models including the cross-sectional means of GT data downloaded for 3 consecutive days (6 consecutive occasions), with and without applying 2x12 moving average, and 3 augmented models including the cross-sectional means of GT data downloaded for 7 consecutive days (14 consecutive occasions), with and without applying 2x12 moving average.

Data in the period 2008–2016 were used as training set, while data for year 2017 were used as test set. The Mean Absolute Percentage Error (MAPE) was considered as a measure of forecast error. Relative out-of-sample forecast errors were computed as the ratio between the MAPE of each augmented model and the MAPE of the benchmark. As such, a value less than 1 means that a certain augmented model beats the benchmark.

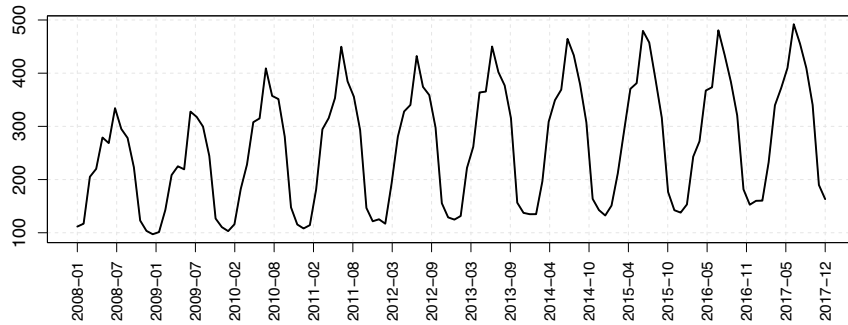


Fig. 1 Official data foreign tourism arrivals in Florence, Italy (thousand persons).

3 Results

All the replications of GT data resulted cointegrated of order 1 with the logarithm of foreign tourist arrivals, according to the Johansen test. Thus, we took first-order differences to estimate the benchmark and the augmented models. The benchmark model resulted an $ARIMA(0, 1, 1)(2, 1, 0)_{12}$ according to an automatic selection

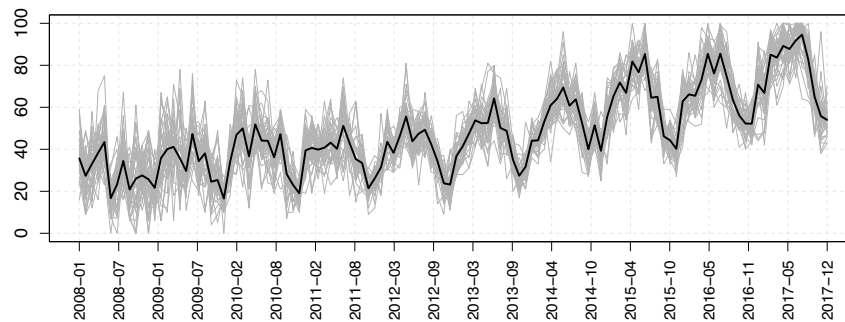


Fig. 2 The downloaded GT data (42 occasions). The bolded series is the cross-sectional mean.

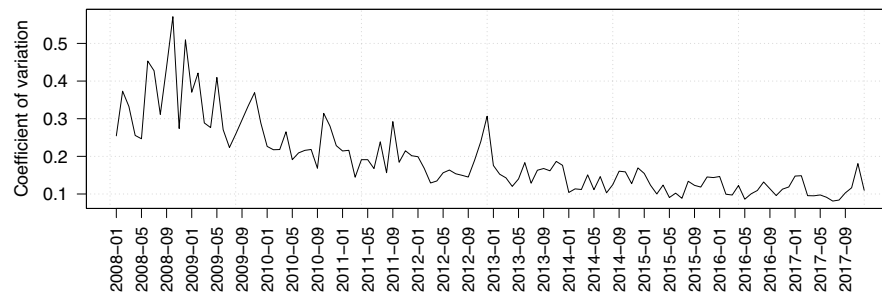


Fig. 3 Cross-sectional coefficient of variation of the downloaded GT data.

procedure based on the Akaike information criterion with correction for small sample size. Relative out-of-sample forecast errors (MAPE) are shown in Table 1.

Results show that not averaged GT data improve forecast 64% of times, to reach 95% of times when 2x12 moving average is applied. Instead, data averaged across 3 and 7 days of download improve forecast 67% and 71% of times, respectively, to reach 100% of times when 2x12 moving average is applied.

4 Discussion

The volume search index provided by Google Trends (GT) have been shown to improve nowcasting of tourism volume by several empirical researches in recent years. However, no effective strategy to deal with the sampling procedure of GT has been proposed, yet.

Table 1 Relative out-of-sample forecast errors (MAPE). A value less than 1 means that a certain augmented model beats the benchmark. The benchmark model is an $ARIMA(0, 1, 1)(2, 1, 0)_{12}$.

Individual data (42 occasions)		
	not filtered	filtered
mean	0.9140	0.6864
minimum	0.2779	0.4023
1st quantile	0.7990	0.6186
median	0.9566	0.6929
3rd quantile	1.0787	0.7407
maximum	1.4278	1.0572
relative errors < 1	64.3%	95.2%
Data average across 3 days of download		
day of download	not filtered	filtered
1 – 3	0.6273	0.5240
4 – 6	0.9614	0.6874
7 – 9	1.0762	0.6675
10 – 12	1.1191	0.5676
13 – 15	0.8940	0.7602
16 – 18	0.9205	0.6143
19 – 21	0.9413	0.6743
relative errors < 1	66.7%	100%
Data average across 7 days of download		
day of download	not filtered	filtered
1 – 7	0.6525	0.5762
8 – 14	1.0130	0.6261
15 – 21	0.9523	0.6500
relative errors < 1	71.4%	100%

By focusing on foreign tourist arrivals in the city of Florence, Italy, we found that the percentage of times where our benchmark forecast is improved significantly increases when averaging data downloaded in different occasions and when filtering each time series using 2x12 moving average. The combination of these two techniques leads to an improvement of forecast in 100%, suggesting that it is not required to replicate the same query several times. Interestingly, moving average alone augments very significantly the percentage of improvement in forecasting, compared to averaging data downloaded in different occasions, suggesting that filtering may be the main determinant for GT data to improve forecast.

Future work may be directed towards the exploration of more complex filtering techniques with the objective to establish whether a small number of replications of the same query, maybe just one, are enough for GT data to improve forecast.

References

1. Andreano, M. S., Benedetti, R., Postiglione, P., and Savio, G. On the use of Google Trend data as covariates in nowcasting: sampling and modeling issues. Proceedings of the Conference of the Italian Statistical Society, Florence, IT, 28-30 June 2017.
2. Athanasopoulos, G., Hyndman, R. J., Song, H., and Wu, D. C. The tourism forecasting competition. *International Journal of Forecast*, **27**: 822–844 (2011).
3. Bisgaard, S., and Kulahci, M. Time series analysis and forecasting by example. John Wiley & Sons, New York, US-NY (2011).
4. Sun, S., Wei, Y., Tsui, K. T., and Wang, S. Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management*, **70**: 1–10 (2019).

Inclusive growth in European countries: a cointegration analysis

La crescita inclusiva nei paesi europei: un'analisi di cointegrazione

Paolo Mariani, Andrea Marletta, Alessandra Michelangeli

Abstract This paper proposes a first measure to detect potential relationships between Gross Domestic Product intended as economic growth indicator and factors related to the job market and inequality measures. This could represent a first signal of a possible economic inclusiveness in some European countries. The timeline covered by data is 1995-2017, while the methodology used is the cointegration analysis. The aim is to establish whether the used variables draw an hypothesis of inclusive dynamic in line with the proposed objectives at European level for 2020.

Abstract Lo scopo di questo paper è proporre una prima misura per eventuali relazioni fra la crescita economica intesa come Prodotto Interno Lordo e alcuni fattori legati al mercato del lavoro e a misure di disuguaglianza ai fini di una possibile inclusività economica in alcuni paesi europei. L'orizzonte temporale copre gli anni dal 1995 al 2017, la metodologia è basata su una analisi di cointegrazione. Il fine è descrivere se le variabili utilizzate disegnano una ipotesi di dinamica inclusiva nel quadro di riferimento degli obiettivi prefissati a livello europeo per il 2020.

Key words: Inclusive growth, cointegration, European countries, GDP

1 Introduction

The economic strategy launched by the European Commission, known as Europe 2020, aims to promote smart, sustainable and inclusive growth. The latter is supposed to foster a high employment economy and enhance social and territorial co-

Paolo Mariani
University of Milano-Bicocca, e-mail: paolo.mariani@unimib.it

Andrea Marletta
University of Milano-Bicocca, e-mail: andrea.marletta@unimib.it

Alessandra Michelangeli
University of Milano-Bicocca, e-mail: alessandra.michelangeli@unimib.it

hesion in the Member States, while reducing the impact of the natural environment. This paper investigates the relationship between economic growth and factors promoting inclusiveness in the EU countries. According to the European Commission, inclusiveness is defined as "How citizens and groups can interact and participate in open policy and decision making". Here we intend inclusiveness as the capability to make young and women participant to the production of national wealth and the decrease of inequality. For this reason we consider as main dimensions of inclusiveness: employment, gender equal opportunities and fair income distribution. A cointegration analysis is used in order to identify long-run economic relationships between economic growth and inclusiveness dimensions in some European countries. The proposed analysis is here intended not to yield considerations about economic policies but only from a statistical point of view. In literature, employment rates and Gini index have been used as indicators of inclusiveness or economic growth in [1] [7].

The rest of the paper is organized as follows. In Section 2 a brief description of data is provided. Section 3 introduces the methodology based on the cointegration analysis. Section 4 presents some preliminary results. Finally, Section 5 is reserved to discussion and final remarks.

2 Data

The aim of the paper is to obtain a first measure to detect potential relationships between economic growth and factors related to the job market and inequality measures. For this reason, the following economic indicators here are used as inclusiveness measures: Gini Index, Employment Rate, Female Employment Rate and the 15-24 years Employment Rate.

Data source for Gross Domestic Product and Employment Rates (total, female and young) is Eurostat. Data about Gini Index are achieved from EU-SILC European Survey on Income and Living Conditions. The 5 time-series indicators show various interval ranges since they are measured in different units. For this reason, it seems to be necessary to use a standardization method to represent them on a single plot. The used method to standardize time series is dividing the indicators for the first value of the time series creating the index numbers.

Data are selected for 31 European Countries from 1995 to 2017. The first condition for analysing the movements of the indicators is to verify the existence of a non-stationary process involving the multiple time series. The stationarity of the time series can be verified using the KPSS (Kwiatkowski Phillips Schmidt Shin) test [6]. In Table 1 the results of the test on some European countries are displayed.

As it is possible to state from Table 1, there is no homogeneity about the time-series stationarity, this leads to a different approach to verify in European countries the presence of inclusive growth.

Here the attention is focused on countries characterized by non-stationarity in GDP and at least another indicator among those selected.

Table 1 KPSS Stationarity test for some European Countries for GDP and inclusive indicators

Country	GDP	Empl.	Gini	Fem.	Young
Austria	No Stat.	Stat.	No Stat.	Stat.	Stat.
Belgium	No Stat.	No Stat.	No Stat.	No Stat.	No Stat.
...
Italy	Stat.	Stat.	No Stat.	Stat.	No Stat.
...
Netherlands	No Stat.	Stat.	Stat.	Stat.	Stat.

In particular, here are reported the results only for Austria and Belgium. Among countries with non-stationary variables, these two have been chosen because their indicators are in line with the average value for Europe [4].

3 Methodology

Since most of the macroeconomic time series follow a random walk process, the attention about the presence of stochastic trends in empirical analysis is grown. Time-series analysts suggest to use variable in difference as a method to obtain stationarity. Moreover, it is not correct to specify a dynamic model only using difference time series, because this leads to no information about the long-run relationships among variables.

A useful technique to detect long-run movements in time series is the co-integration. Variables are defined co-integrated if it exists a stationary linear combination between non-stationary time series. Usually, linear combinations of non-stationary variables are non-stationary too. But sometimes it could happen that a particular linear combination is stationary. If this bond exists, then a long-run movement among the variables of the model is present.

A multivariate time-series $X_t' = [X_{1t}, \dots, X_{kt}]$ is co-integrated if each single component is a time-series $I(1)$ and it exists at least a weighted vector $a' = [a_1, \dots, a_k]$ such as the linear combination $a'X_t$ is stationary.

The a vector, if exists, is named co-integration vector. If $k = 2$, it is possible to obtain a unique co-integration vector normalizing to 1 the first component of a . If $k > 2$, it could exist $h < k$ linearly independent vectors, usually collected in the co-integration matrix A .

The co-integration has a clear meaning in Economics. An example of co-integrated system is the consumption model proposed by Davidson, Hendry, Srba e Yeo (1978) [2]. Their results suggest that time-series of consumption and income have unit-roots, but in the long-run the ratio consumption/income is constant, so that the difference between the logarithm of consumption and the logarithm of income is a stationary process.

When $k = 2$ and the rank of the co-integration matrix $r = 1$, the best way to achieve the long-run relationship is the procedure suggested by Engle-Granger [3]. Let be x_t and y_t two time-series integrated of lag 1, the OLS regression model is:

$$X_t = \hat{\gamma} + \hat{\delta}Y_t + \hat{v}_t$$

where \hat{v}_t are the residuals of the model.

When $k > 2$ and the rank of the co-integration matrix $r > 1$, a recursive technique is used to compute the rank of the cointegration matrix, the Johansen test based on hypothesis tests on different values assigned to h in H_0 .

Starting from the VAR model written in Error Correction Model, the null hypothesis H_0 : is the presence of h co-integration relationships versus the alternative hypothesis H_1 : the presence of k co-integration relationships. The test statistics is obtained applying the transformed $-2\log(\Lambda)$ on the ratio Λ between likelihood under H_0 and H_1 .

4 Application and results for Austria and Belgium

As briefly introduced in the previous section, the two countries studied in this work are Austria and Belgium. From results in table 1, it is possible to note that they denote two different scenarios.

For Austria, it is possible to hypothesize the presence of a relationship only involving 2 variables, GDP and Gini Index, so it is possible to use an Engle-Granger procedure in order to understand the causality relationship. The estimation of a VAR(1) for Austria returns significant only lagged GDP_{t-1} coefficient for GDP_t and both lagged GDP_{t-1} and $Gini_{t-1}$ for $Gini_t$ (see p-value in brackets). This could suggest an influence only of GDP on Gini Index.

$$GDP_t = 0.927GDP_{t-1}(0.000) + 0.107Gini_{t-1}(0.112)$$

$$Gini_t = 0.267GDP_{t-1}(0.007) + 0.669Gini_{t-1}(0.000)$$

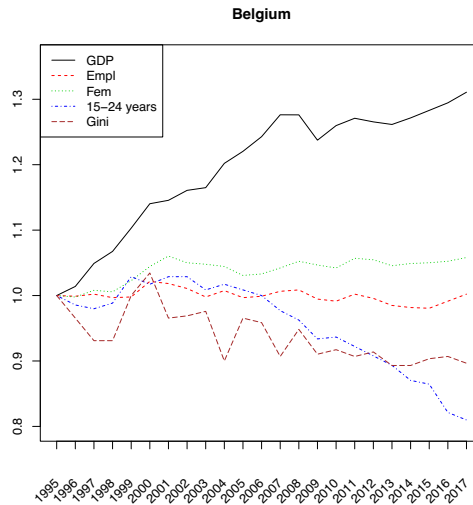
The Engle-Granger procedure detects the direction using a F test comparing the unrestricted model in which y is explained by the lags of y and x the restricted model in which y is only explained by the lags of y . Since only p-value of the model in which $Gini_t$ is the dependent variable is significant (see p-value in brackets), this is a confirmation that GDP has an influence on Gini Index and not vice versa.

$$GDP \text{ --- } > Gini \quad F = 0.224(0.641)$$

$$Gini \text{ --- } > GDP \quad F = 10.848(0.004)$$

About Belgium, it is necessary a deeper analysis searching for that linear combination of time series able to give back a stationary process underlining a long-run balance among indicators. Firstly, in figure 1, we displayed the normalised time-series. From a graphical analysis, GDP seems to be growing, the Gini index and the young employment rate to be decreasing and total and female employment rate appears to be in a border-line situation (stationarity or slightly growth). KPSS test suggested in table 1 to keep all variables in the model for the moment.

Fig. 1 GDP, employment rates and Gini Index, for Belgium, 1995-2017



Once verified the non-stationarity of the time series, it is possible to hypothesize the estimate of a VAR (Vector AutoRegressive) Model. The estimate of the lag order p has been obtained using the typical loss functions that suggested to use $p = 1$, so a VAR(1) is estimated. The presence of a unit-root in estimated VAR(1) for Belgium underlines the non-stationarity of the process. This means that a co-integration among indicators is plausible.

The Johansen Test for co-integration states the rank of co-integration matrix $r = 1$, hypothesizing the existence of one long-run relationships among variables. To identify this relationships, a Vector Error Correction Model (VECM) with 1 long-run component has been estimated producing only these significant coefficients:

- $GDP_t = Empl_{t-1}, Gini_{t-1}, EC1$
- $Empl$ no coefficients
- $Gini = GDP_{t-1}, EC1$
- $Fem, Empl$ no coefficients
- $Young$ no coefficients

where $EC1$ is the coefficient identifying the long-run component.

Even in this scenario, it is plausible to hypothesize the existence of a co-movement when considering both the GDP and the Gini index. From this exploratory analysis for Austria and Belgium, the growth of the GDP looks like to be inclusive only influenced by Gini index and not by the employment rates.

Just basing our considerations about these two countries, it could seem that the employment did not represent an incentive for the inclusive growth, while a supposed link between GDP and Gini index is plausible. Despite the presence of a

growth and a decrease in inequality, there is no influence neither on total employment and the young or female employment.

5 Conclusion and future researches

The article proposes an analysis in detecting co-movements for economic indicators related to the inclusive growth using a co-integration approach. Data from Eurostat measured the growth using the Gross Domestic Product and the inclusiveness through 3 employment rates (total, female and young) and the Gini index as measure of fair income distribution for the period 1995-2017.

Results on some European countries showed a spread heterogeneity in analysing these relationships. Austria and Belgium presented a possible long-run co-movement involving GDP and Gini index. This could be a first index of the existence of a connection between the presence of a growth and a decrease in inequality. On the other hand, from data evidence, it is not confirmed the engagement of the employment measures in this process of inclusive growth.

Future research will focus the attention on results considering all European countries trying to hypothesize the presence of clusters of areas using a spatial correlation matrix. Another interesting development of the research might inspect the presence of structural shocks using dummy variables related to some particular events as the introduction of the Euro currency in 2002 or the financial crisis of 2009.

References

1. Bhalla, S. Inclusive growth? Focus on employment. *Social Scientist*, 24-43. (2007)
2. Davidson, J. E., Hendry, D. F., Srba, F., Yeo, S. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *The Economic Journal*, 88(352), 661-692. (1978)
3. Engle, R. F., Granger, C. W. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251-276. (1987)
4. Hoj, J. Enhancing the Inclusiveness of the Labour Market in Belgium, OECD Economics Department Working Papers, No. 1009, OECD Publishing, Paris, <https://doi.org/10.1787/5k4dl080l5d2-en>. (2013)
5. Johansen, S., Juselius, K. Maximum Likelihood Estimation and Inference on Cointegration – with Applications to the Demand for Money, *Oxford Bulletin of Economics and Statistics*, 52, 2, 169–210. (1990)
6. Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., Shin Y. Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root. *Journal of Econometrics* 54, 159–178. (1992)
7. Zhuang, J., Ali, I. Poverty, inequality, and inclusive growth in Asia. *Poverty, Inequality, and Inclusive Growth: Measurement, Policy Issues, and Country Studies*, 1-32. (2010)

ESCO- the European Labour Language: a conceptual and operational asset in support of labour governance in complex environments

ESCO il linguaggio europeo del lavoro: uno strumento concettuale ed operativo per le politiche del lavoro in contesti complessi

Cristilla Martelli, Laura Grassini, Adham Kahlawi, Maria Flora Salvatori, Lucia Buzzigoli

Abstract Labour market policies and decision processes need to rely on knowledge systems able to represent modern labour market complex dynamics. This contribute is oriented to discuss the role that ESCO standard labour language may play in this perspective

Abstract *Le politiche e i processi decisionali che coinvolgono il mercato del lavoro devono poter poggiare su sistemi di conoscenza in grado di rappresentarne fedelmente le complesse dinamiche evolutive. Questo contributo intende riflettere sul ruolo che il linguaggio europeo ESCO può ricoprire per raggiungere questi obiettivi.*

Key words: labour market, labour semantic models, network analysis.

1 Introduction

Labour market is a dynamic and complex system, which involves multiple stakeholders, mutually interrelated: individuals, families, policies makers, training, and education systems, economic stakeholders and investors. All these actors, may have different visions and requirements, and need support in their decisional processes [4].

Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze.
Corresponding author: Cristina Martelli, e-mail: cristina.martelli@unifi.it

In this perspective, it is a main objective to gather intelligence and knowledge able to represent this complexity, able to support along with the long lead times often implied in the management of these policy decisions.

A main rational for this knowledge system is to avoid labour market failures and to ensure a better fit among skills supply and requests, to improve community performances and ensure a better life for individuals and their families.

To pursue these challenges, the European labour language ESCO represents a main asset as it keeps together a high descriptive granularity with the precision of a standard language, in the context of a well designed conceptual model described and manageable with semantic web utilities.

This contribution is organised as follows: after having introduced ESCO, in its main characteristic and potentialities in a complex network perspective (§ 1), an exploration of its network characteristics will be performed (§2). Some main suggestions will be discussed in the concluding remarks (§3), where a reflection on the adoption of ESCO language in the context of the service oriented to skill supply and request is done in the perspective of the generation of high quality administrative data to be exploited in a statistical reuse perspective.

2 The European labour language ESCO, as a complex conceptual network

ESCO is the multilingual classification of European, Skills, Competences, Qualifications and Occupations. ESCO is an important deliverable to support the Europe 2020 strategy and the New Skills Agenda for Europe [6]: it covers three different domains (the three pillars of ESCO: (i) occupations; (ii) knowledges/skills; (iii) qualifications. Each one of these pillars contains concepts and terms that help to describe supply, needs and offers on the labour market in a standardized machine readable manner using Linked Open Data [5].

ESCO, as a language, is particularly suitable to respect the complex dimension of labour market and it allows to model and support knowledge system able to describe and support such complexity.

In this contribution, complexity is used in its strict meaning: according to Barabási conceptualization and language [1, 3], a complex network is not just a matter of interconnected agents: complex systems display several organizing principles, which are, at some levels, encoded in their topology.

In the following a recap of complex network main principles in the perspective of labour market description.

Small world property, it means that, despite their size, in most networks there is a relatively short path between any two nodes. The distance between two nodes is defined as the number of edges along the shortest connecting path: the most popular small world expression [10] is the “six degrees of separation” concept, proposed in 1967 by the social psychologist Stanley Milgram. *Small world property* has inspired

Contribution Title

sociologists and politicians and it has often been declined in terms of interdependency and multiculturalism:

In this contribution, we re-think small world property as an asset in support of professional reconversion between the actual occupation and the one to which one wants to access.

Clustering. Another common property of complex networks is an inherent tendency to clustering, quantified by the clustering coefficient [13] or by the transitivity coefficient [9]: in most real networks, the clustering coefficient is typically much larger than it is in a random network of an equal number of nodes and edges.

The existence of clusters of occupations or skills makes the labour market more readable and allows the identification of priorities in training and policies

Degree distribution. Not all nodes in a network have the same number of edges. Usually, in a random graph, as the edges are placed randomly, the majority of nodes have approximately the same degree and their distribution $P(k)$, which gives rise to the probability that a randomly selected node has exactly k edges, is a Poisson distribution. Empirical results show that for complex networks, the degree distribution significantly deviates from a Poisson distribution. Such networks are called scale-free [3]. A scale free network originates when a node, having to establish a new association, prefers to link itself with a node already characterized by many connections, contributing to an exponential growth of its connections inside the network: in this way the functional form of $P(k)$ deviates from the Poisson distribution of $P(k)$ expected for a random graph.

The scale free property is an indicator of the existence of the evolutive characteristics of an occupation, which is able to “attract” new skills and knowledges.

3 Exploring ESCO network

The adherence of ESCO to the complexity paradigm provides important keys to labor market interpretation. In the following, some network analysis exercises have been performed to explore features, characteristics and potentialities resulting from the adoption of the ESCO standard.

ESCO classification comprises 2941 occupations and more than 10500 skills (many “transversal skills”), with the average number of skills per occupation greater than 20. Table 1 shows the six most required skills and the six most skilled occupations.

Table 1 Top 6 most required skills (left) and top 6 most skilled occupations (right)

Skill	N. occupations	Occupation	N. skills
manage staff	254	chiropractor	136
create solutions to problems	240	music therapist	98
quality standards	239	clinical psychologist	95
have computer literacy	236	biomedical scientist	90
troubleshoot	236	nurse responsible for general care	90
adhere to organisational guidelines	233	physiotherapist	90

A first exploration of ESCO is carried out through network analysis, which was already used by other authors (for example [11]). The analysis is accomplished with *sna* R package.

We can state that if occupation (node) A and occupation B do have at least h common skills, then there is a link between A and B. The value h works as a threshold for activating the connection between paired occupations. In fact, as the number of skills per occupation is high, a threshold $h=1$ produces an iperconnected graph, with a density greater than 0.10. A choice $h=5$ gives a density of 0.011, an average degree of 32.3, a transitivity coefficient of 0.808, and produces 325 isolates (Figure 1).

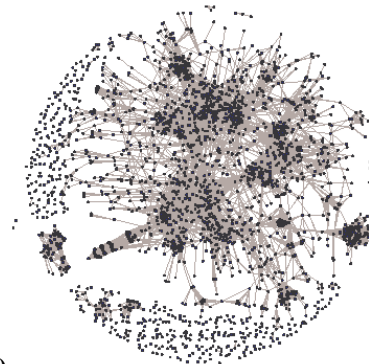
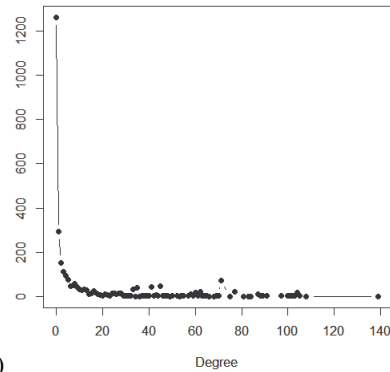


Figure 1 Network of occupations ($h=5$)

The choice of other threshold values of h does not produce a network with characteristics similar to the three prototypes cited above: with increasing h a larger number of isolates have been produced (1200 with $h=10$; Figure 2).

Figure 2 Frequency of node degree ($h=10$)

Another exploration for an in depth knowledge of skill distributions can be done by the association rules technique. An association rule has two parts: an antecedent (if) and a consequent (then) where a consequent is an item frequently occurring in combination with the antecedent. This technique can help in finding frequent associated skills and, therefore, a more rational pruning of skills. Table 2 reports the top ten association rules of the method by fixing the support to 0.03 and the maximum size of the itemset to 4. The analysis is achieved with *arules* R package.

Table 3 Top 6 association rules

Lhs (antecedent)	rhs	lift
{apply teaching strategies, curriculum objectives, prepare lesson content}	{apply intercultural teaching strategies}	27.679
{curriculum objectives, guarantee students' safety, prepare lesson content}	{apply intercultural teaching strategies}	27.670
{assess students, curriculum objectives, prepare lesson content}	{apply intercultural teaching strategies}	27.642
{give constructive feedback, perform classroom management}	{prepare lesson content}	27.486
{apply intercultural teaching strategies, curriculum objectives}	{prepare lesson content}	27.486
{apply teaching strategies, give constructive feedback, perform classroom management}	{prepare lesson content}	27.486

3 Conclusions

The availability of a multilingual classification of occupations together with skills, competences and qualifications like ESCO can be of great help in organizing the knowledge on the European labour market and also on the sector of education and training, and therefore can improve both the matching between qualifications and labour market needs (first level) and the matching between jobseekers and employers (second level) [7, 8, 12].

The matching at the first level is based on the pillar of qualifications, which is still in progress in ESCO, and is based on widely used classifications like the FoET for education (Fields of Education and Training from the International Standard Classification of Education, version 2013) and ISCO2008 for occupations (International Standard Classification of Occupations). It could produce important effects in planning education and training paths according to labour market evolution.

At the second level, the availability of a system that breaks down CV items of jobseekers into standardized knowledge, skills and competencies facilitates the matching with recruiters that share the same common language. The European jobs network EURES uses ESCO taxonomy for the exchange of vacancies and CVs and foreshadows a standardized exchange of information between Public Employment Services by means of syntactic interoperability that could be a valuable instrument to translate administrative data into statistical ones. The section dedicated to the integration of ESCO with the process of implementation of the national system of skills contained in the strategic plan of ANPAL (the Italian National Agency for Active Labour Market Policies) can be interpreted in this framework.

References

1. Albert, R., Barabási, A.L., Statistical mechanics of complex networks, *Reviews of modern physics*, 74.1, 47 (2002).
2. ANPAL, Piano strategico triennale (2018).
3. Barabási, A.L.; Albert R. Emergence of scaling in random networks. *Science*, 509-512 (1999).
4. CEDEFOP, ETF, ILO Developing skills foresights, scenarios and forecasts. *Guide to anticipating and matching skills and jobs*, Vol. 2. (2016).
5. De Smedt, J., le Vrang, M., Papantoniou, A. ESCO: Towards a Semantic Web for the European Labor Market. Linked data on the web (LDOW), Florence (2015).
6. ESCOmedia:
https://ec.europa.eu/esco/portal/escopedia/European_Skills_44_Competences_44_Qualifications_and_Occupations_40_ESCO_41
7. European Commission. Three tools to facilitate online job matching throughout Europe, Luxembourg (2010).
8. European Commission, Proposal for a Regulation of the European Parliament and of the Council on a European network of Employment Services, workers' access to mobility services and the further integration of labour markets, Commission Staff Working Document 9, Brussels (2014).
9. Luke, D. A. A User's Guide to Network Analysis in R, Springer (2015)
10. Milgram S. The small world problem. *Psychology today*, 2.1, 60-67 (1967).
11. Öpik, R., Kirt, T., Liiv, I. Megatrend and Intervention Impact Analyser for Jobs: a visualization method for labour market intelligence. *Journal of Official Statistics*, 34(4), 961-979 (2018)
12. Siekmann, G. & Fowler, C. Identifying work skills: international approaches, NCVER, Adelaide (2017).
13. Watts D. J.; Strogatz S. H. Collective dynamics of 'small-world' networks. *Nature*, 440-442 (1998).

Hidden Markov Models for High Dimensional Data

Hidden Markov Models per dati ad alta dimensionalità

Martino, A., Guatteri, G., Paganoni, A.M.

Abstract Hidden Markov Models (HMMs) are a very popular tool used in many fields to model time series data. In the last years, big data have become more and more important. For this reason, we extend the HMMs to the case of high dimensional data. Specifically, we focus on the case of functional data, by taking into consideration a sequence of multivariate curves that evolves in time. The functional observations are linked to the state of the HMM according to a similarity function, which depends on some metric in Hilbert spaces. We first assess our results in a simulation study and then apply our model to a case study regarding the weather in Canada.

Abstract *Gli Hidden Markov Model (HMM) sono uno strumento molto utilizzato in molti campi per modellizzare serie temporali di dati. Negli ultimi anni, i big data sono diventati sempre più importanti. Per questa ragione, estenderemo gli HMM al caso di dati ad alta dimensionalità. In particolare, ci concentriamo sul caso di dati funzionali, prendendo in considerazione una sequenza di curve multivariate che evolve nel tempo. Le osservazioni funzionali sono collegate allo stato dell'HMM in base ad una funzione di similarità che dipende da una metrica in spazi di Hilbert. Dopo aver valutato i risultati in uno studio di simulazione, applichiamo il nostro modello ad un caso studio meteorologico.*

Key words: Hidden Markov Models, Functional Data

Martino Andrea, Guatteri Giuseppina, Paganoni Anna Maria
Department of Mathematics, Politecnico di Milano, via Bonardi 9, 20133, Milan, Italy,
e-mail: andrea.martino@polimi.it
e-mail: giuseppina.guatteri@polimi.it
e-mail: anna.paganoni@polimi.it

1 Introduction

Hidden Markov Models (HMMs) are a popular method for modeling time series. They consist of a Markov model in which the underlying states visited by the Markov process are unobservable (i.e. hidden) but the distribution that generates the output depends on the state [3]. We only consider models where the state space of the hidden variables is discrete. Usually, HMMs consider an univariate or multivariate output. In this work, we want to extend the use of HMMs to functional observations. In Section 2 we present the model, adding some information about the theory of HMMs. In Section 3 we present a simulation study to assess the performance of the model while in Section 4 we see a case study application to a dataset regarding canadian weather. All the analysis have been carried out using the statistical software R [2].

2 The model

The aim of this work is to develop a proper Hidden Markov Model (HMM) in the multivariate functional framework. Typically, HMMs are used to model univariate or multivariate data taking values in \mathbb{R}^d , $d \geq 1$. Let us consider a multivariate random curve $\mathbf{X} = \{\mathbf{X}(t)\}_{t \in I} = \{X_1(t), \dots, X_J(t)\}_{t \in I}$, with $J \geq 1$ and I compact interval of \mathbb{R} , as a random element of $(L^2(I))^J$ equipped with the Borel σ -algebra, such that $\{X_j(t)\}_{t \in I} \in (L^2(I))$ for any $j \in \{1, \dots, J\}$.

We can define a Hidden Markov Model [1] as a process $\{(Q_k, \{\mathbf{X}_k(t)\}_{t \in I})\}_{k \geq 0}$ on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\{\mathbf{X}_k(t)\}_{t \in I}$ is a multivariate random curve and $\{Q_k\}_{k \geq 0}$ is a Markov chain with a discrete and finite state space $\{s_1, \dots, s_N\}$, with $N \geq 1$, transition matrix $A = \{a_{ij}\} = \mathbb{P}(Q_k = s_j | Q_{k-1} = s_i)$ and initial distribution \mathbf{v} , where $v_i = \mathbb{P}(Q_0 = s_i)$. Given the process $\{Q_k\}_{k \geq 0}$, $\{\{\mathbf{X}_k(t)\}_{t \in I}\}_{k \geq 0}$ is a sequence of conditionally independent multivariate functions and $\{\mathbf{X}_k(t)\}_{t \in I}$ only depends on Q_k for each k . We denote the emission function of \mathbf{X}_k conditionally on the event $\{Q_k = s_i\}$ with $b_i(\cdot; \boldsymbol{\mu}_i)$, for any $i = 1, \dots, N$, where $\boldsymbol{\mu}_i$ is a functional representative of state s_i ; specifically, $b_i(\cdot; \boldsymbol{\mu}_i)$ represents the likelihood that the function \mathbf{X}_k is emitted from state s_i . We can completely define our HMM with the set of parameters $\lambda = (\mathbf{v}, A, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N)$.

In this work, we use distances between functions to construct the emission functions $b_i(\cdot; \boldsymbol{\mu}_i)$, $i = 1, \dots, N$. Let us denote with d a generic distance in $(L^2(I))^J$; the likelihood that a realization of the stochastic process \mathbf{X} is emitted from state s_i can be written as

$$b_i(\cdot; \boldsymbol{\mu}_i) = h\left(d(\cdot, \boldsymbol{\mu}_i)\right)$$

where $h : \mathbb{R} \rightarrow \mathbb{R}$ is a function that transforms the distance into a similarity measure. In particular, we will use the L^2 distance that, in the multivariate functional framework, is defined as follows:

$$d_{L^2}(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\| = \sqrt{\sum_{l=1}^J \int_I (a_l(t) - b_l(t))^2 dt}.$$

An important step of our algorithm is the initialization. Since we want our algorithm to converge and to be as much precise as possible, we perform a functional k -means algorithm [6] on the dataset of curves to find the initial centroids.

Let us denote with \mathbf{x} an output sequence of observation functions of the HMM and with $\mathcal{L}(\lambda|\mathbf{x})$ the likelihood function of all the parameters of the model given \mathbf{x} . In the literature of HMMs, there are usually three problems to tackle (see, e.g., [7] and [3]): find $\mathcal{L}(\lambda|\mathbf{x})$ for the realization \mathbf{x} , find the set of parameters $\lambda^* = \arg\max_{\lambda} \mathcal{L}(\lambda|\mathbf{x})$, find the best state sequence $Q = (Q_1, \dots, Q_K)$ that explains \mathbf{x} , given \mathbf{x} and λ . As usually done in the literature, to address these problems we use the forward-backward procedure, the Baum-Welch algorithm and the Viterbi algorithm, respectively.

To fully describe our algorithms, we need to introduce two further quantities. We define $\xi_k(i, j) = \mathbb{P}(Q_k = s_i, Q_{k+1} = s_j \mid X_1 = x_1, \dots, X_k = x_k, \lambda)$, the probability of being in state s_i at time k , and state s_j at time $k+1$, given the model and the observations and $\gamma_k(i) = \mathbb{P}(Q_k = s_i \mid X_1 = x_1, \dots, X_k = x_k, \lambda) = \sum_{j=1}^N \xi_k(i, j)$, the probability of being in the state s_i at time k , given the observations and the model. From [7], we can write the log-likelihood of our model as

$$\begin{aligned} \log(\mathcal{L}(\lambda|\mathbf{x})) = & \underbrace{\sum_{j=1}^N \gamma_1(j) \log v_j}_{\text{term 1}} + \underbrace{\sum_{i=1}^N \sum_{j=1}^N \left(\sum_{k=2}^K \xi_k(i, j) \right) \log a_{ij}}_{\text{term 2}} \\ & + \underbrace{\sum_{j=1}^N \sum_{k=1}^K \gamma_k(j) \log b_i(\mathbf{x}_k; \boldsymbol{\mu}_j)}_{\text{term 3}}. \end{aligned} \quad (1)$$

Using this expression, we can apply an EM algorithm to compute all the parameters of the HMM. Even though until now our formulas only consider a single observation sequence, they can be extended to the more general case of multiple observations. Let us denote the set of L observation sequences as $\mathcal{X} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)})$, where $\mathbf{x}^{(l)} = (\mathbf{x}_1^{(l)}, \mathbf{x}_2^{(l)}, \dots, \mathbf{x}_{K_l}^{(l)})$ is the l -th observation sequence of length K_l . We assume all the sequences to be independent from each other; our goal is to adjust the parameters of the model λ to maximize the likelihood $\mathcal{L}(\lambda|\mathcal{X})$. The term we want to maximize, in the setting of multiple sequences, becomes

$$\text{term 3} = \sum_{l=1}^L \sum_{k=1}^{K_l} \sum_{j=1}^N \gamma_k(j) \log b_i(\mathbf{x}_k; \boldsymbol{\mu}_j). \quad (2)$$

Since we want to estimate the functional representatives $(\mu_{lj})_{i=1, \dots, N; j=1, \dots, J}$ of the states of the HMM, we compute for every state i and every component j

$$\hat{\mu}_{ij} = \underset{\mu_{ij}}{\operatorname{argmax}} \sum_{l=1}^L \sum_{k=1}^{K_l} \gamma_k(j) \log b_i(x_{lk}; \mu_{ij}).$$

To perform this step, we extend all the estimators commonly used in the functional data framework into the theory of functional HMM.

3 Simulation Studies

We generate three samples of length n of i.i.d. realizations for three independent bivariate stochastic processes $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ in $(L^2(I))^J$, with $J = 2$. Each sample is emitted from a different state of a 3-state HMM having the following parameters:

- **State 1:** $v_1 = 1, a_{11} = 0.6, a_{12} = 0.3, a_{13} = 0.1, \mathbf{m}_1(t) = \begin{pmatrix} t(1-t) \\ 2t \end{pmatrix};$
- **State 2:** $v_2 = 0, a_{21} = 0.1, a_{22} = 0.8, a_{23} = 0.1, \mathbf{m}_2(t) = \begin{pmatrix} t^2(1-t) \\ t^2 \end{pmatrix};$
- **State 3:** $v_3 = 0, a_{31} = 0, a_{32} = 0, a_{33} = 1, \mathbf{m}_3(t) = \begin{pmatrix} t(1-t)^2 \\ \frac{1}{2}t^3 \end{pmatrix}.$

where $\mathbf{v} = (v_i)$ is the vector of the initial probabilities of the state, $A = (a_{ij})$ is the transition matrix and $\mathbf{m}_i(t), i = 1, \dots, N$, represent the real means of each sample. For each state, the sample is generated using the same exponential covariance kernel $C(s, t) = ae^{-b|s-t|}$, $a = 0.1, b = 0.3$. The first problem consists in the choice of the number of states, since it is a priori unknown. We begin by running our algorithm for $N = 2, \dots, 5$ number of states and by computing each time the AIC and BIC criteria. Since both criteria reach the minimum value for $N = 3$ states, we choose this value as the "optimal" number of states for the HMM. After choosing the number of states, we summarize our results along 100 repetitions of our algorithm to estimate the parameters of the HMM. To have a better understanding of the results, we compute the mean square error (MSE) and the standard deviation (SD) of the parameters and we show the obtained results in Tab. 1. As we can see, all the parameters are very well estimated, both in terms of mean and standard deviation of the parameters.

Moreover, we can obtain some further information about the clustering structure of our data. Specifically, we use our model and apply the Viterbi algorithm on the output obtained from the Baum-Welch algorithm, to estimate the best state sequence and compare it with the output of the k -means algorithm, based on the same distance. In particular, in Tab. 2 we compare the Correct Classification Rate (CCR) of the number of curves obtained by applying both methods along with the MSE of the functional representatives of a state or cluster $m_i, i = 1, 2, 3$, i.e. the values of the distances between the real and the estimated means. By comparing the results, we see obvious advantages of our method, since the CCR is higher and all the distances are smaller. We can conclude that, not only our method is able to detect the temporal structure behind the sequences of functional data and estimate all the parameters of

Parameter	MSE (SD)
a_{11}	$3.71 \cdot 10^{-2}$ ($9.23 \cdot 10^{-3}$)
a_{12}	$8.30 \cdot 10^{-3}$ ($1.17 \cdot 10^{-2}$)
a_{13}	$8.01 \cdot 10^{-2}$ ($4.10 \cdot 10^{-2}$)
a_{21}	$2.89 \cdot 10^{-2}$ ($2.32 \cdot 10^{-3}$)
a_{22}	$2.07 \cdot 10^{-3}$ ($1.70 \cdot 10^{-3}$)
a_{23}	$9.72 \cdot 10^{-4}$ ($1.69 \cdot 10^{-3}$)
a_{31}	$1.31 \cdot 10^{-3}$ ($9.29 \cdot 10^{-3}$)
a_{32}	$7.10 \cdot 10^{-8}$ ($3.54 \cdot 10^{-7}$)
a_{33}	$1.32 \cdot 10^{-3}$ ($9.31 \cdot 10^{-3}$)
v_1	$2.00 \cdot 10^{-2}$ ($1.41 \cdot 10^{-1}$)
v_2	$2.00 \cdot 10^{-2}$ ($1.41 \cdot 10^{-1}$)
v_3	$< 2 \cdot 10^{-16}$ ($< 2 \cdot 10^{-16}$)

Table 1 MSE (SD) of the HMM parameters for 100 simulation runs of the Baum-Welch algorithm with $N = 3$ states for the HMM.

	Viterbi Algorithm	k -means algorithm
CCR	0.857	0.591
m_1	0.085	0.131
m_2	0.636	0.890
m_3	0.953	1.051

Table 2 C.C.R. and distances between the real and estimates means for 100 simulation runs of the Viterbi and k -means algorithm.

the underlying hidden states but, applying the Viterbi algorithm, we can also cluster the curves obtaining good values of accuracy.

4 Case Study

In this last part, we apply the described model to a real dataset regarding the weather in Canada [4] and [5]. In particular, our data consists of temperature and precipitation at 35 different locations in Canada, averaged over 1960 to 1994. The stations where the functional measurements take place are the statistical units and they change over the months, so we have 12 measurements for each station. We run the models for $N = 2, \dots, 5$ states and we obtain the best value for $N = 4$. By applying our Baum-Welch algorithm, we obtain the following results:

$$A = \begin{pmatrix} 0.66 & 0.00 & 0.00 & 0.34 \\ 0.00 & 0.43 & 0.29 & 0.28 \\ 0.00 & 0.22 & 0.78 & 0.00 \\ 0.15 & 0.35 & 0.00 & 0.50 \end{pmatrix}$$

We associated the two extreme states in terms of both temperature and precipitation

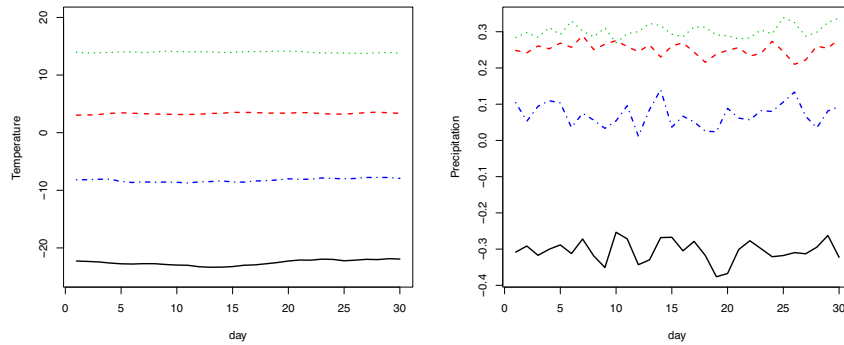


Fig. 1 Plots of the functional representatives for the 4 states of temperature and precipitation.

to summer (green lines, State 3) and winter (black lines, State 1) while the other ones represent middle seasons. We cannot make a clear distinction among the seasons because the position of the station influences a lot the analysis. In particular, since we know the location of each station, we can notice that the stations closer to the ocean never reach State 1 while the ones in the Arctic region never reach State 3. Moreover, by looking at matrix A , we see how the probabilities of staying in a state are usually higher and that there is a clear distinction between State 1 and 3, since you can never go from one state to the other.

References

1. Cappé, O., Moulines, E., & Rydén, T. (2009). Inference in hidden markov models. Springer Publishing Company, Incorporated
2. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
3. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286.
4. Ramsay J., Wickham H., Spencer Graves and Giles Hooker (2018). fda: Functional Data Analysis. R package version 2.4.8. <https://CRAN.R-project.org/package=fda>
5. Ramsay J., Silverman B. W. (2002) Applied functional data analysis - Methods and case studies, Springer Series in Statistics. Springer-Verlag, New York.
6. Tarpey T., Kinateder K. K. (2003) Clustering functional data. *J. Classification*, 20:93–114, 2003.
7. Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). Hidden Markov models for time series: an introduction using R. Chapman and Hall/CRC.

Classification of Italian classes via bivariate semi parametric multilevel models

Classificazione delle classi italiane per mezzo di modelli bivariati a effetti misti semi parametrici

Chiara Masci, Francesca Ieva, Tommaso Agasisti and Anna Maria Paganoni

Abstract In this work, we propose a bivariate semi-parametric mixed-effects model where the random effects are assumed to follow a discrete distribution with an unknown number of support points, together with an Expectation-Maximization algorithm to estimate its parameters - the BSPeM algorithm. This model for hierarchical data enables the identification of subpopulations within the higher level of hierarchy. The bivariate setting allows to estimate the distributions of the model coefficients related to each response variable as well as their joint distribution. In the case study, we apply the BSPeM algorithm to data about Italian middle schools, considering students nested within classes, and we identify subpopulations of classes that have different effects on reading and mathematics student achievements. The strength of the proposed bivariate semi-parametric mixed-effects model is twofold: first, it is an innovative model that can be applied in many classification problems dealing with multiple outcomes; second, when applied in the educational context, given the strong connections among student learning processes in different school subjects, it results to be extremely informative in modeling the correlation between multiple class effects. The identified subpopulations of classes are then explained in terms of teacher characteristics, teaching practices and class body composition.

Abstract *In questo lavoro, proponiamo un modello a effetti misti semi parametrico bivariato, in cui gli effetti casuali seguono una distribuzione discreta con un numero di masse non noto a priori, insieme ad un algoritmo EM per la stima dei suoi parametri - algoritmo BSPeM. Questo modello permette di identificare la pre-*

Chiara Masci
Politecnico di Milano, via Bonardi 9, Milan e-mail: chiara.masci@polimi.it

Francesca Ieva
Politecnico di Milano, via Bonardi 9, Milan e-mail: francesca.ieva@polimi.it

Tommaso Agasisti
Politecnico di Milano, via Lambruschini 4/a, Milan e-mail: tommaso.agasisti@polimi.it

Anna Maria Paganoni
Politecnico di Milano, via Bonardi 9, Milan e-mail: anna.paganoni@polimi.it

senza di sotto-popolazioni all'interno del più alto livello della gerarchia di dati con struttura gerarchica. L'utilizzo di un modello bivaraito permette di stimare sia le distribuzioni dei coefficienti del modello relativi a entrambe le risposte sia la loro distribuzione congiunta. Nel caso studio, applichiamo l'algoritmo BSPEM ai dati INVALSI (Istituto Nazione per la valutazione del sistema educativo di istruzione e di formazione), considerando gli studenti annidati dentro le classi, e identifichiamo sotto-popolazioni di classi che hanno effetti diversi sul rendimento degli studenti in italiano e matematica. I vantaggi del modello proposto sono due: primo, è un modello innovativo che può essere applicato in vari problemi di classificazione con output multipli; secondo, quando viene applicato nel contesto educativo, date le forti interazioni tra i processi di apprendimento degli studenti in diverse materie, risulta essere estremamente utile per stimare la correlazione tra i vari effetti classe. Le sotto-popolazioni di classi identificate sono poi caratterizzate in termini di caratteristiche degli insegnati, metodi di insegnamento e composizione delle classi.

Key words: EM algorithm, multivariate statistics, semi-parametric mixed-effects models, student achievements.

1 Introduction

The educational system is a hierarchical system in which different levels of grouping are nested within each others: students are nested within classes, that are in turn nested within schools, that are in turn nested within districts and so on so forth. Each one of these levels has a role in the student learning process. Measuring how much of the variability in student education is due to each grouping level is not easy, but, it is essential for evaluating the role of education. A broad and rich literature about *school value-added* based on test scores, intended as the difference in test performance of students in a school and the average performance of schools populated by students with a comparable level of prior achievement (and other student characteristics) [12, 13], has been developed in the last decades. Even though the measurement of school value-added continuously receives attention, decades of educational effectiveness confirm that differences between pupils is more within schools than between them [5, 11]. In this perspective, the concept of school value-added can be transferred to the class level, speaking about class value-added. Class peers, class climate and, especially, teachers considerably affect the student learning process.

In this work, exploiting previous results in the research about school and class value-added in the Italian education context, we propose a study that is innovative both from a methodological and an interpretative point of view. First of all, we develop a bivariate, i.e. for a bivariate response variable, semi-parametric mixed-effects linear model. Secondly, we show how this new method can be effective in the research about class effectiveness, by applying it in a case study that faces the new issue of the identification of clusters of Italian classes, standing on their joint effect on student achievement trends in reading and mathematics.

The model that we propose is a bivariate two-level linear model where the coefficients of random effects, under semi-parametric assumptions, follow a bivariate discrete distribution with an unknown number of mass points. Each group (observation of the second level of grouping) is assigned to a bivariate subpopulation of groups, that is represented by specific values of the parameters of the bivariate mixed-effects linear model. The distribution of the coefficients of random effects is a bivariate discrete distribution where each dimension is allowed to have a different finite number, unknown *a priori*, of mass points. This formulation permits to estimate the marginal distribution of the random effects related to each one of the two response variables and, moreover, to estimate the joint distribution of random effects related to the two response variables, investigating the correlation among them. In this perspective, we do not create a full ranking of the highest level effects, but instead we identify subpopulations of effects and we attribute each group to a single subpopulation. Together with the model, we propose an Expectation-Maximization (EM) algorithm to estimate its parameters (BSPEM algorithm).

The proposed methodology is the multivariate extension of the semi-parametric mixed-effects linear model proposed in [6], that is totally new to the literature. In particular, the model presented in [6] on which we base our multivariate model enters in the research line about the identification of subpopulations of the Growth Mixture Models (GMM) [8, 10] and of Latent Class Mixture Models (LCMM) [7], but with the novelty and the advantage that, contrarily to these existing methods, it does not need to fix *a priori* the number of latent subpopulations to be identified. Moreover, numerous extensions and applications of GMM and LCMM has been done [2, 9], but none of them include the modeling of a multivariate answer variable, where the latent subpopulations structure of groups (higher level of hierarchy) are allowed to differ across the responses, i.e. are response-specific.

In the application to the INVALSI data, the two-levels model, in which we consider students as first level and classes as second one, aims at identifying a latent clustering structure of classes where, within each cluster, the effect of the classes on their student achievement trends across years are similar.

2 The Dataset

The INVALSI database [1] contains information about 18,242 students attending the third year of junior secondary school in the year 2016/2017, nested within 1,082 classes. At pupil's level, we consider reading and mathematics INVALSI test scores at grade 8 (RS and MS); reading and mathematics INVALSI test scores at grade 5, three years before, of the same students; the socio-economic index (ESCS), the gender and the immigrant status of students. Moreover, INVALSI in the survey 2016/2017, by means of teacher questionnaires, collected information about teachers characteristics (age, education, gender...), teaching practices, class-body composition and geographical area.

3 Methodology

We consider the case of a bivariate semi-parametric two-level model with P fixed covariates, one random intercept and one random covariate. The model takes the following form:

$$\begin{aligned} \mathbf{Y}_i &= \begin{pmatrix} y_{1,i} \\ y_{2,i} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \mathbf{X}_i + \begin{pmatrix} \mathbf{b}_{1,i} \\ \mathbf{b}_{2,i} \end{pmatrix} \mathbf{Z}_i + \boldsymbol{\varepsilon}_i \quad i = 1, \dots, N \\ \boldsymbol{\varepsilon}_i &= \begin{pmatrix} \varepsilon_{1,i} \\ \varepsilon_{2,i} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \quad ind. \end{aligned} \quad (1)$$

where, in our application, N is the total number of classes; \mathbf{Y}_i is the $(n_i \times 2)$ -dimensional matrix of student achievements in reading and mathematics at grade 8 in class i ; \mathbf{X}_i is the $(n_i \times P)$ -dimensional matrix of ESCS, gender and immigrant status of students in class i ; \mathbf{Z}_i is the $(n_i \times 2)$ -dimensional matrix of the same students achievements in reading and mathematics at grade 5 (three years before) in class i . We use these variables at student level because we are interesting in modeling the association between student achievements at grade 5 and 8, across different classes and in different school subjects, adjusting the model for the other variables that, based on the literature, may influence this association, that are the ESCS, the gender and the immigrant status. $\mathbf{b} \in R^4$ is the vector containing the coefficients of random effects. \mathbf{b} follows a discrete distribution P^* with $M \times K$ support points, where M and K are not known *a priori*. P^* can then be interpreted as the mixing distribution that generates the density of the stochastic model in (1). The ML estimator \hat{P}^* of P^* can be obtained following the theory of mixture likelihoods in [3] and [4], where the author proves the existence, discreteness and uniqueness of the non-parametric maximum likelihood estimator of a mixing distribution, in the case of exponential family densities. The ML estimator of the random effects distribution can be expressed as a set of points $(\mathbf{b}_{11}, \dots, \mathbf{b}_{MK})$, where $M \leq N$, $K \leq N$ and $\mathbf{b}_{mk} \in R^4$ for $m = 1, \dots, M$, $k = 1, \dots, K$ and a set of weights (w_{11}, \dots, w_{MK}) , where $\sum_{m=1}^M \sum_{k=1}^K w_{mk} = 1$ and $w_{mk} \geq 0$ for each $m = 1, \dots, M$ and $k = 1, \dots, K$. Given this, we develop an algorithm for the joint estimation of $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}$, (b_{11}, \dots, b_{MK}) and (w_{11}, \dots, w_{MK}) , that is performed through the maximization of the likelihood, mixture by the discrete distribution of the random effects,

$$\begin{aligned} L(\boldsymbol{\beta}, \mathbf{c}_{mk}, \boldsymbol{\Sigma} | \mathbf{y}) &= \sum_{m=1}^M \sum_{k=1}^K \frac{w_{mk}}{\sqrt{|\det(2\pi\boldsymbol{\Sigma})|^J}} \times \\ &\times \exp \left\{ \sum_{i=1}^N \sum_{j=1}^{n_i} -\frac{1}{2} \left(y_{1,ij} - b_{1,1m} - \sum_{p=1}^P \beta_{1p} x_{1p,ij} - b_{1,2m} z_{1,ij} \right)^T \boldsymbol{\Sigma}^{-1} \right. \\ &\quad \left. \left(y_{2,ij} - b_{2,1k} - \sum_{p=1}^P \beta_{2p} x_{2p,ij} - b_{2,2k} z_{2,ij} \right) \right\} \end{aligned} \quad (2)$$

with respect to Σ , β and $(\mathbf{b}_{mk}, w_{mk})$, for $m = 1, \dots, M$ and $k = 1, \dots, K$. Each class i , for $i = 1, \dots, N$ is therefore assigned to a cluster mk , for $m = 1, \dots, M$ and $k = 1, \dots, K$. $J = \sum_{i=1}^N n_i$. The EM algorithm is an iterative algorithm that alternates two steps: the expectation step (E step) in which we compute the conditional expectation of the likelihood function with respect to the random effects, given the observations and the parameters computed in the previous iteration; and the maximization step (M step) in which we maximize the conditional expectation of the likelihood function. Moreover, given N starting support points, during the iterations of the EM algorithm, we reduce the support of the discrete distribution standing on both two criteria: the former is that we fix a threshold D and if two points are closer than D they collapse in a unique point; the latter is that we remove points with very low weights (choosing a given threshold \tilde{w}) and that are not associated to any class. When two points \mathbf{b}_{l*} and \mathbf{b}_{m*} (if considering the coefficients related to the first response) collapse to a unique point, because their Euclidean distance is smaller than D , we obtain a new mass point $\mathbf{b}_{(lm)*} = \frac{\mathbf{b}_{l*} + \mathbf{b}_{m*}}{2}$ with weight $w_{(lm)*} = w_{l*} + w_{m*}$. The thresholds D and \tilde{w} are two complexity parameters that affect the estimation of the semi parametric distribution: the higher is D , the lower is the number of clusters. Notice that the numbers of support points M and K are computed by the algorithm and we do not have to fix it a priori. Since we do not have to specify a priori the number of support points, the semi parametric mixed-effects model could be interpreted as an unsupervised clustering tool for nested data.

4 Results

The BSPM algorithm applied to INVALSI data identifies 5 subpopulations for the class effects in mathematics and 4 subpopulations for reading. Estimates of the parameters are shown in Table 1.

The estimated matrix of the weights W and the variance/covariance matrix Σ of the errors are the following:

$$\hat{W} = \begin{pmatrix} 0.0000 & 0.0007 & 0.0003 & 0.4571 \\ 0.0054 & 0.0518 & 0.0047 & 0.3220 \\ 0.0022 & 0.0000 & 0.0023 & 0.0204 \\ 0.0068 & 0.0312 & 0.0082 & 0.0179 \\ 0.0043 & 0.0111 & 0.0029 & 0.0507 \end{pmatrix} \quad \hat{\Sigma} = \begin{pmatrix} 0.455 & 0.183 \\ 0.183 & 0.451 \end{pmatrix}. \quad (3)$$

The different subpopulations are characterized by different class effects and the distribution of the classes over the 20 mass points is not homogeneous, suggesting the presence of a significant correlation among the subpopulations distributions related to the two response variables. The characterization of the subpopulations *a posteriori* reveals that subpopulations differ in terms of teaching practices and geographical distribution.

First response variable						
	$\hat{c}_{1,1}$ (intercept)	$\hat{c}_{1,2}$ (math5)	\hat{w}_1 (weight)	$\hat{\beta}_{11}$ (ESCS)	$\hat{\beta}_{12}$ (gender)	$\hat{\beta}_{13}$ (immigrant)
m=1	0.295	0.719	0.458			
m=2	-0.181	0.463	0.384			
m=3	0.762	0.463	0.025	0.089	-0.055	0.048
m=4	-1.301	0.112	0.064			
m=5	0.366	0.291	0.069			

Second response variable						
	$\hat{c}_{2,1}$ (intercept)	$\hat{c}_{2,2}$ (read5)	\hat{w}_2 (weight)	$\hat{\beta}_{21}$ (ESCS)	$\hat{\beta}_{22}$ (gender)	$\hat{\beta}_{23}$ (immigrant)
k=1	-2.848	-0.101	0.019			
k=2	-0.622	0.262	0.095			
k=3	-1.556	0.188	0.018	0.095	0.219	-0.083
k=4	0.054	0.544	0.868			

Table 1 Estimates of the coefficients of Eq. (1) obtained by the BSPM algorithm.

References

1. INVALSI website
<http://www.invalsi.it/>
2. L. I. Lin et al.: Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine*, 19(2):255-270, 2000.
3. B. Lindsay: The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, 11(1), 86-94, 1983.
4. B. Lindsay: The geometry of mixture likelihoods, part II: the exponential family. *The Annals of Statistics*, 11(3), 783-792, 1983.
5. C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni: Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. *Journal of Applied Statistics*, 44(7):1296-1317, 2017.
6. C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni: Semi-parametric mixed-effects models for the unsupervised classification of Italian schools. *Journal of the Royal Statistical Society - Series A*. In press.
7. C. McCulloch, H. Lin, E. Slate, and B. Turnbull: Discovering subpopulation structure with latent class mixed models. *Statistics in medicine*, 21(3):417-429, 2002.
8. B. Muth  n; Latent variable analysis. *The Sage handbook of quantitative methodology for the social sciences*, 345:368, 2004.
9. B. Muth  n and T. Asparouhov: Growth mixture modeling with non-normal distributions. *Statistics in Medicine*, 34(6):1041-1058, 2015.
10. B. Muth  n and K. Shedden: Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, 55(2):463-469, 1999.
11. T. Perry: English value-added measures: Examining the limitations of school performance measurement. *British Educational Research Journal*, 42(6):1056-1080, 2016.
12. S. W. Raudenbush and J. Willms: The estimation of school effects. *Journal of educational and behavioral statistics*, 20(4):307-335, 1995.
13. J. D. Willms and S. W. Raudenbush: A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of educational measurement*, 26(3):209-232, 1989.

Data Mining Application to Healthcare Fraud Detection: Two-Step Unsupervised Clustering Method for Outlier Detection with Administrative Databases

Data Mining Applicato al Riconoscimento Frodi in Sanita': Algoritmo a Due Step per l'Identificazione di Outliers con Database Amministrativi

Massi Michela C., Ieva Francesca, Lettieri Emanuele

Abstract This study aims at exploiting Administrative Databases to identify potentially fraudulent providers. It focuses on the DRG upcoding practice, i.e. the tendency of coding within Hospital Discharge Charts (HDC), codes for provided services and inpatients health status so to make the hospitalization fall within a more remunerative DRG class. The model here proposed is constituted by two steps: one first step entails the clustering of providers, in order to spot outliers within groups of similar peers; in the second step, a cross-validation is performed, to verify the suspiciousness of the identified outliers. The proposed model was tested on a database relative to HDC collected by Regione Lombardia (Italy) in a time period of three years (2013-2015), focusing on the treatment of heart failure.

Massi Michela Carlotta
MOX - Modellistica e Calcolo Scientifico
Dipartimento di Matematica
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy

Center for Analysis Decisions and Society
Human Technopole
Palazzo Italia, Via Cristina Belgioioso, 28, 20157 Milano, Italy
e-mail: michelacarlotta.massi@polimi.it

Ieva Francesca
MOX - Modellistica e Calcolo Scientifico
Dipartimento di Matematica
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
e-mail: francesca.ieva@polimi.it

Lettieri Emanuele
Department of Management, Economics and Industrial Engineering
Politecnico di Milano
Lambruschini Street 4/c, 20100 Milano, Italy
Tel.: +39-02-23994077
e-mail: emanuele.lettieri@polimi.it

Abstract *Questo studio ha lo scopo di sfruttare i Database Amministrativi per identificare operatori sanitari potenzialmente fraudolenti. Si focalizza sulla pratica dell'upcoding, i.e. la tendenza a registrare sulle Schede di Dimissione Ordinaria, codici relativi ai servizi prestati e allo stato di salute del paziente, in modo da far ricadere il ricovero in una classe DRG pi remunerativa. Il modello proposto costituito da due step: il primo riguarda il clustering degli ospedali, per identificare outliers tra gruppi di operatori simili; il secondo punta a validare la sospettosit degli outlier identificati. Questo modello stato testato su un database amministrativo relativo al trattamento di Scompenso Cardiaco, collezionato da Regione Lombardia in un periodo di tre anni (2013-2015).*

Key words: Data Mining, Healthcare Fraud, DRG Upcoding, Administrative Database

1 Introduction

Fraud in the context of Healthcare has different perpetrators (e.g. hospitals, medical figures, private facilities) and different dynamics.

The *upcoding* practice consists in classifying a patient in a DRG, or registering treatment codes (in case of Fee for Service¹ payment systems) that produce higher reimbursements [Simborg (1981)]. This practice by public hospitals is more likely to be due to unintentional errors by coders, or misunderstandings with doctors; while, when talking about private hospitals or medical practices, it could actually be due to profit maximising purposes [Silverman (2004), O'Malley (2005)].

Given that Healthcare is the target of large public and private investments (on average, in OECD countries, 15% of the government budget is allocated for this purpose), independently from the geographical and political setting, this sector is a rather appetible one for frauds. In this ever-growing healthcare industry, using manual countermeasures to fight frauds is not enough. Given the ever-growing availability of digital data, the adoption of data mining techniques might help to reach better results and more efficient processes, in terms of both time and costs. Developing tailored algorithms would allow for the restriction of the pool of investigated providers, including those acting cautiously and perpetrating fraud within the limits of those indicators monitored according to general policies [Musal (2010)].

In this paper, we deal with the development of a novel systematic and quantitative approach to fraud detection, with a focus on the upcoding practice, because of the extremely relevant economic impact this kind of fraud has on the system [Steinbusch (2007)]. The objective is to propose a novel tool to support human decisions in the preliminary phases of screening providers to spot suspects eligible for a more in depth investigation, including those with a more cautious approach to fraud. Because of the large availability of Administrative Databases, we decided to exploit this type of data as our source of information. All the analysis described in

¹ A method in which doctors and other health care providers are paid for each service performed. Examples of services include tests and office visits. [Healthcare.gov]

the following Sections were performed using R [R Core].

The proof of concept of our method, which is generally adaptive to any kind of system, was developed in the Italian context, Lombardy Region in particular, studying the *behavior* of hospitals by exploiting regional administrative data. This region represents a relevant benchmark in the Italian healthcare landscape and as such it was deemed an interesting starting point for our testing, but the flexibility of the developed method allows for its generalizability to other contexts.

2 Literature Review

Data mining approaches to healthcare fraud detection are still at their infancy, but recently gaining momentum. Even though limited, some contributions to the literature exist. We can group the existing methodologies of fraud detection into three main groups: supervised, unsupervised, or hybrids of the two [Aral (2012)]. Supervised techniques [Pearson (2006), Kumar (2010), Francis (2011), Kirlidog (2012)] despite their undeniable potential and predictive power, exhibit the risk of focusing on old patterns and losing predictive capability as new records are evaluated over time [Joudaki (2015)]. For this reason, unsupervised techniques are the most adopted [Yang (2006), Shan (2008), Shan (2009), Luo (2010), Musal (2010), Tang (2011), Konijin (2012), Shin (2012), Liu (2013), Konijin (2013), Konijin (2015), Bauder (2016), Capelleveen (2016)], tackling the intrinsic characteristic of fraud of changing in time, according to the arising regulations or control systems. Hybrid techniques or on-line processing systems take the best of both approaches [Aral (2012), Ngufor (2013), Kose (2015)], creating a combination of supervised and unsupervised methods.

As previously mentioned, this paper focuses on the upcoding practice, that can be declined in several ways and has drawn the attention of most data mining researches because of the extremely high impact it has on healthcare expenditures. The majority of the available literature attempts to spot providers with very high claiming episodes, which distinguish themselves as *evident* outliers - one clear example in [Capelleveen (2016)]. However, the evolving nature of modern fraud, has pushed some researchers to change direction and try and identify providers that have a more cautious approach to fraud, which would be neglected in the aforementioned high-claiming groups. The *behavioral* models suggested by Musal *et al.* [Musal (2010)] and Shin *et al.* [Shin (2012)] try indeed to respond to this objective. However, those techniques have still very limited application, even though their usefulness in reducing the overall level of fraud in the system is undeniable. This lack of interest can be justified by the risk of these models to exhibit higher false positive rates, and the lower amount of recovery from each correctly spotted fraudster [Musal (2010)]. However, as reported in [Shin (2012)], types of fraud are growing increasingly sophisticated, and patterns detected from fraudulent and non-fraudulent behaviors become rapidly obsolete because of rapid changes in behavior, and fraudulent providers are becoming smarter in finding more cautious approaches

which usually prevent them from being investigated [Bolton (2002)].

In conclusion, from the study of previous contributions emerges the need for more data mining methods capable to adapt to changing behaviors of fraudsters, and identifying more cautious approaches to fraud. For this reason, the algorithm proposed in this study belongs to this more complex and less tackled pool of unsupervised *behavioral* methodologies.

3 Methods

The proposed model was tested on an Administrative Database collected by Lombardy Region, composed by Hospital Discharge Charts (HDCs) from 2013 to 2015. Patients' data were at first anonymized, and hospital confidentiality was preserved. To improve the model's performance, one specific disease was selected to search for anomalous behaviors by hospitals. Therefore, an extraction process was performed as first step, focusing on Heart Failure cases.

Since this research was performed in an unsupervised setting (ie. no indication about fraudulent records was available at the beginning of the process), and no interaction with experts was planned, all choices were driven by insights gathered from the Literature Review. At first, three different datasets were built: the Hospital Discharge Charts (HDCs) Dataset, with one HDC per row, the Hospital Dataset, aggregating data with hospitals as statistical unit, and the Patients Dataset, where each observation is a patient treated at least once by the providers. From literature, a list of variables deemed interesting to evaluate fraudulent behaviors was collected (e.g. [Berta (2010), Ekin (2017)], or the Upcoding Index in [Silverman (2004)]). Such variables have subsequently been extracted or reproduced to be added to the available datasets.

The algorithm proposed in this research is constituted by two fundamental steps:

One *First Step*, in which the Hospitals Dataset is exploited. This dataset was composed by descriptive variables (i.e. number of patients treated by the hospital, average cost, average length of stay, degree of specialization for treatment of HF, etc.), and integrated with the values of r_{ij} [Ekin (2017)], which represents the ratio between the probability that the HF-related DRG i is registered by provider j and the probability that DRG i is registered in all the population of N hospitals, calculated for every HF-related DRG registered by the provider. The vector of r_{ij} represents, for each provider, a description of the '*behavior*' of that provider w.r.t. the treatment of HF.

Grouping hospitals on the basis of their characteristics and their behavior in the treatment of HF, makes it possible to distinguish homogeneous groups of providers, allowing us to spot the providers *behaving* differently from their similar peers. For this reason, the algorithm here proposed entails a k-mean clustering on the aforementioned hospitals' information, with the aim of identifying clusters of similar providers.

Once the clusters have been identified, the Euclidean distance of each element of the

Table 1 Variables adopted for cross-validation with hospitals' characteristics and casemix

Variable	Dataset	Note
Specialization	Hospital	Number of HF specific cases
Perc. DRGs with CC	Hospital	Complexity of the casemix
Upcoding Index	Hospital	Berta <i>et al.</i> , 2010[Berta (2010)]
Average Cost	Hospital	Expensiveness of the casemix
Age	Patients	Patient complexity indicator
Length of Stay	Patients, HDC	Patient complexity indicator
Comorbidity	Patients, HDC	Patient complexity indicator, estimated as in [Gagne (2011)]
Total Costs	Patients, HDC	Cost of care per Patient
Cost / Length of Stay	HDC	Cost of care against quantity of care needed
Cost / Comorbidity	Hospital	Cost of care against quantity of care needed

cluster by the corresponding centroid is computed. Figure 2 shows such distribution for all the hospitals of our dataset. Given such distribution, it is possible to define a threshold (say the 95th quantile), where hospitals having a distance greater than this threshold are labelled as 'providers needing further investigation'.

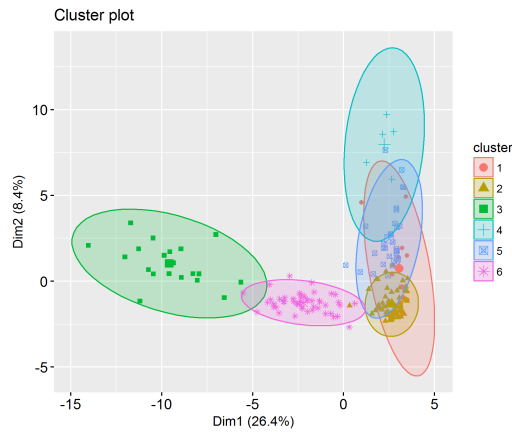
A *Second Step* is then performed by the algorithm. Since *behavioral methods* demonstrate to have higher false positive rates - even though stronger in identifying cautious fraudsters [Musal (2010)] - a further validation of results is recommended. This phase is indeed useful to controllers to create a visual dashboard to support an informed skimming of the suspects identified in the first phase. In this way, their attention and in-depth investigations will be focused on a restricted number of cases, with a higher probability of detecting actual frauds. This passage entails the use of variables describing Hospitals deemed interesting in literature to identify fraudulent providers, together with indexes estimating the complexity of the casemix of patients they faced. Both groups of variables are useful to try and verify whether such outliers could be considered suspects w.r.t. upcoding fraud or, on the opposite, justified in their particular behavior because of the complex patients' population they treated. In Table 1 are listed the variables adopted for the cross-validation.

4 Results

From the application of the clustering algorithm on the data at hand - after the estimation of the optimal number of clusters with the WSS method - resulted the six clusters of hospitals represented in Figure 1.

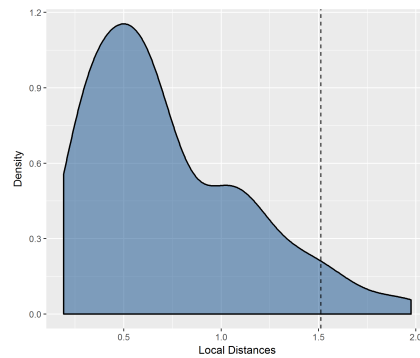
The number of clusters was chosen on the basis of a robustness analysis where the different values of Silhouette obtained for different k values were compared to identify the optimal clusters' configuration. In addition, evaluating clusters characteristics, clear differences among clusters were recognized w.r.t. some dimensions of interest: for this reason, the result was deemed satisfactory.

Fig. 1 Clusters of hospitals resulting from *k-mean* clustering. The clusters are plotted against the first two Principal Components resulting from a PCA.



Once the clusters were generated and all distances calculated, 10 hospitals resulted as outliers adopting the 95th percentile threshold.

Fig. 2 Local distances distribution of hospitals from the center of their cluster. The 95th percentile threshold is highlighted by the vertical dotted line.



Among these, 8 were private providers, and 2 public.

The complete analysis can be found in [Massi (2018)], but one illustrative case could be that of a private provider with very low level of specialization, and a very high Upcoding Index, that when cross-validated with its casemix of patients confirmed its suspect behavior (e.g. both Comorbidity Score and average LOS were low, but the reimbursements requested per hospitalization were higher than the average of providers).

This is just an example of how this procedure could constitute an important method to support performances' monitoring in the Healthcare domain. Indeed, combining the first and the second step of this method we get a twofold goal: first, we provide a flexible tool for screening a huge amount of data which are currently available (e.g. Administrative Databases), exploiting their use for performance assessment; moreover we limit the number of false positive to be deeply investigated by controllers.

References

- [Aral (2012)] Aral K.D., Güvenir H.A., Sabuncuoğlu I, Akar A.R., (2012) A Prescription Fraud Detection Model, *Comput. Methods Prog. Biomed.*, **106** (1), pp. 37–46.
- [Bauder (2016)] Bauder R., Khoshgoftaar T.M., Seliya N., (2017) A survey on the state of healthcare upcoding fraud analysis and detection, *Health Services and Outcomes Research Methodology*, **17** (1), pp. 31–55.
- [Berta (2010)] Berta P., G. Callea, G. Martini, G. Vittadini, (2010) The effects of upcoding, cream skinning and readmissions on the Italian hospitals efficiency: A population-based investigation, *Economic Modelling*, **27**, pp. 812–821.
- [Bolton (2002)] Bolton R.J., Hand D.J., (2002) Statistical Fraud Detection: A Review, *Statistical Science*, **17** (3), pp. 235–255.
- [Capelleveen (2016)] Van Capelleveen G., Mannes P., Roland M., Dallas T., Van Hillegersberg J., (2016) Outlier detection in healthcare fraud: A case study in the Medicaid dental domain, *International Journal of Accounting Information Systems*, **21**, pp. 1831.
- [Ekin (2017)] Ekin T., Ieva F., Ruggeri F., Soyer R., (2017) On the Use of the Concentration Function in Medical Fraud Assessment, *The American Statistician*, **71** (3), pp. 236–241.
- [Francis (2011)] Francis C., Pepper N., and Strong H., (2011), Using support vector machines to detect medical fraud and abuse 2011 *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 8291–8294.
- [Gagne (2011)] Gagne J.J., Glynn R.J., Avorn J., Levin R., Schneeweiss S., (2011) A combined comorbidity score predicted mortality in elderly patients better than existing scores, *Journal of clinical epidemiology*, **64** (7), pp. 749–759.
- [Healthcare.gov] Healthcare.gov, (2018), [online] <https://www.healthcare.gov/glossary/fee-for-service/>
- [Joudaki (2015)] Joudaki H., Rashidian A. *et al.*, (2015) Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature, *Global Journal of Health Science*, **101**, **7**, p. 378.
- [Kirlidog (2012)] Kirlidog M., Cuneyt A., (2012) A Fraud Detection Approach with Data Mining in Health Insurance, *Procedia - Social and Behavioral Sciences*, **62**, pp. 989–994.
- [Konijin (2012)] Konijin R.M., Kowalczyk W., (2012) Hunting for Fraudsters in Random Forests, *Hybrid Artificial Intelligent Systems: 7th International Conference - Proceedings*, **Part 1**, pp. 174–185.
- [Konijin (2013)] Konijin R.M., Duivesteijn W., Kowalczyk W., Knobbe A., (2013) Discovering Local Subgroups, with an Application to Fraud Detection, *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference - Proceedings*, **Part 1**, pp. 1–12.
- [Konijin (2015)] Konijin R.M., Duivesteijn W., Meeng M., Knobbe A., (2015) Cost-based quality measures in subgroup discovery, *Journal of Intelligent Information Systems*, **45** (3), pp. 337–355.
- [Kose (2015)] Kose I., Gokturk M., Kilic K., (2015) An Interactive Machine-learning-based Electronic Fraud and Abuse Detection System in Healthcare Insurance, *Appl. Soft Comput.*, **36** (C), pp. 283–299.
- [Kumar (2010)] Kumar M., Ghani R., Mei ZS, (2010) Data Mining to Predict and Prevent Errors in Health Insurance Claims Processing, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 65–74.
- [Liu (2013)] Liu Q., Vasarhelyi M., (2013) Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information, *29th World Continuous Auditing and Reporting Symposium*, **Part 1**.
- [Luo (2010)] Luo W., Gallagher M., (2010) Unsupervised DRG Upcoding Detection in Healthcare Databases, *2010 IEEE International Conference on Data Mining Workshops*, pp. 600–605.
- [Massi (2018)] Massi M.C., Ieva F., Lettieri E., (2018) Data Mining Application to Healthcare Fraud Detection: A Two-Step Unsupervised Clustering Model for Outlier Detection with Administrative Databases, *Mox Report*, 49/2018

- [Musal (2010)] Musal R.M., (2010) Two models to investigate Medicare fraud within unsupervised databases, *Expert Systems with Applications*, **37** (12), pp. 8628–8633.
- [Ngufor (2013)] Ngufor C., Woytusiak J., (2013) Unsupervised labeling of data for supervised learning and its application to medical claims prediction, *Computer Science*, **14** (2).
- [O'Malley (2005)] O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM., (2005) Measuring diagnoses: ICD code accuracy., *Journal of Health Economics*, **40** (5), pp. 1620–1639.
- [Pearson (2006)] Pearson R.A., Murray W., Mettenmeyer T., (2006) Finding anomalies in Medicare, *Electronic Journal of Health Informatics*, **1** (1), p. 2.
- [R Core] R Development Core Team (2009), R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. [Online] <http://www.R-project.org>
- [Silverman (2004)] Silverman E., Skinner J, Medicare upcoding and hospital ownership, *Journal of Health Economics*, **23** (2), pp. 369–389 (2004)
- [Simborg (1981)] Simborg DW, (1981) DRG creep: a new hospital-acquired disease, *New Engl J Med*, **304**(26), pp. 1602-1604.
- [Shan (2008)] Shan Y, Jeacocke D., Murray D.W., Sutinen A., (2008) Mining Medical Specialist Billing Patterns for Health Service Management, *Proceedings of the 7th Australasian Data Mining Conference*, **87**, pp. 105–110.
- [Shan (2009)] Shan Y., Murray D.W., Sutinen A., (2009) Discovering Inappropriate Billings with Local Density Based Outlier Detection Method, *Proceedings of the Eighth Australasian Data Mining Conference*, **101**, pp. 93–98.
- [Shin (2012)] Shin H., Park H., Lee J., Jhee W.C., (2012) A Scoring Model to Detect Abusive Billing Patterns in Health Insurance Claims, *Expert Systems with Applications*, **39** (8), pp. 7441–7450.
- [Steinbusch (2007)] Steinbusch P., Oostenbrink B., Zuurbier J., Schaepekens F., (2007) The Risk of Upcoding in Casemix Systems: A Comparative Study, *Health Policy*, **81**, pp. 298–299.
- [Tang (2011)] Tang M., Mendis B., Sumudu U., Murray D.W., Hu Y., Sutinen A., (2011) Unsupervised Fraud Detection in Medicare Australia, *Proceedings of the Ninth Australasian Data Mining Conference*, **121**, pp. 103–110.
- [Yang (2006)] Yang WS, Hwang SY, (2006) A process-mining framework for the detection of healthcare fraud and abuse, *Expert Systems with Applications*, **31** (1), pp. 56–68.

Multivariate analysis and biodiversity partitioning of a demersal fish community: an application to Lazio coast

Analisi multivariata e partizione della biodiversità di una comunità di specie demersali: un'applicazione alla costa laziale

M. Mingione, G. Jona Lasinio, S. Martino, F. Colloca

Abstract Community ecology aims at understanding which factors determine the composition and the dynamics of species assemblages at different spatio-temporal scales. A large amount of the existing literature on this topic analyzes species individually or in a small number. However, it is well-established that species interact with each other and these interactions influence the assembly of a community. In this work, we study the composition of the demersal community occurring in front of the Lazio coast. First, we estimate biodiversity components measured by Tsallis entropy in order to describe possible diversity variations in the available time window (1995-2017). Then, we run the Parafac procedure to describe the community at species level.

Abstract Lo studio delle comunità biotiche si pone l'obiettivo di capire quali siano i fattori che determinano la composizione e le dinamiche di assemblaggi di specie a diversi livelli di scala spazio-temporale. Una larga parte della letteratura scientifica su questo argomento tratta le specie singolarmente o in piccoli gruppi. E' noto che le specie interagiscono tra di loro e queste interazioni influiscono nella struttura e funzioni delle comunità. In questo lavoro, è stata studiata la composizione della comunità delle specie demersali del Mar Tirreno centrale (costa del Lazio). La prima parte del lavoro stima le componenti della biodiversità, utilizzando l'entropia di

Marco Mingione

University of Rome "La Sapienza", Statistical Science Department, Piazzale Aldo Moro 5, 00185, e-mail: marco.mingione@uniroma1.it

Giovanna Jona Lasinio

University of Rome "La Sapienza", Statistical Science Department, Piazzale Aldo Moro 5, 00185, e-mail: giovanna.jonalasinio@uniroma1.it

Sara Martino

Norwegian University of Science and Technology, NO-7491 Trondheim, Norway, e-mail: sara.martino@ntnu.no

Francesco Colloca

Institute for Marine Biological Resources and Biotechnologies - CNR, Mazara del Vallo (TP) e-mail: francesco.colloca@iamc.cnr.it

Tsallis, per descrivere le eventuali variazioni della biodiversità nel periodo 1995-2017. In seguito, l'analisi multivariata (Parafac) è stata utilizzata per descrivere la struttura della comunità.

Key words: Biodiversity partitioning, Parafac, demersal fish community

1 Introduction

Understanding the distribution and abundance of species is a primary goal of ecological research[10]. In this respect, several attempts have been made in modeling species distribution, also including in the model regressors that may possibly affect the abundance of species. These models can be used to illuminate prevalence, predict biodiversity and richness, quantify species turnover, and assess response to climate change [9, 5, 3, 7, 11]. A further key objective of these models is interpolation, i.e. to predict species response at locations that have not been sampled.

In this first effort, we are interested in simply describing how a heavily exploited Mediterranean community of fishes and shellfishes has changed in the last 25 years. To this aim, we used the Tsallis entropy to describe the evolution at different spatial scales and the Parafac analysis to describe the structure of the community.

2 The data

We analyze data coming from the MEDITS programme, which aims at conducting co-ordinated bottom trawl surveys in the Mediterranean Sea¹. In particular, we consider 40 trawl stations located along central Tyrrhenian Sea (Lazio coast), see Figure 1, sampled from 1995 to 2017. A total of 323 different demersal species were recorded², including selaceans, teleosteans, cephalopods and crustaceans. The response variable is the relative abundance index of each single species expressed as Kg/Km^2 for each trawl station and year. Thus, the dataset has $323 \times 40 \times 23 \simeq 30 \times 10^5$ rows. The trawl stations are georeferenced, longitude, latitude and depth (x, y, z) are therefore available. We use the latter to classify trawl stations as follows: (i) *Continental shelf stations* if $z \in [0m, 130m)$; (ii) *Stations at the shelf-edge* if $z \in [130m, 200m)$; (iii) *Continental slope stations* if $z \in [200m, 750m)$. The location of the trawl stations and their classification are shown in Figure 1. There are 14 trawl stations at the shelf, only 8 trawl stations at the edge and 19 trawl station at the slope.

¹ <http://www.sibm.it/SITO%20MEDITS/principaleprogramme.htm>

² at least once in the considered time window

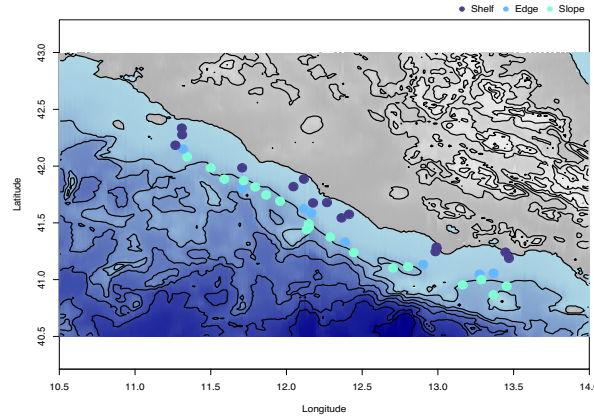


Fig. 1 Location of the fishing station in the study area, coloured by depth class.

2.1 Biodiversity partitioning

In ecology, γ -diversity is a common indicator used to estimate the diversity (or entropy) of a community [12]. The main idea is that the diversity can be decomposed into two terms: the diversity at a local scale (*within* the single station), called α -diversity and the diversity *between* the local scales, called β -diversity. These two components are assumed to be independent, thus the γ -diversity can be expressed as $\gamma = \alpha \times \beta$ (*Biodiversity Partitioning*). The term diversity can be used alternatively to the term entropy, the latter being a monotone transformation of the former. Tsallis generalized entropy, proposed by [8], allows to estimate biodiversity indeces in a unified framework. Given a discrete set of species probabilities $p = \{p_s\}$ and any real number q , the Tsallis entropy of order q is defined as:

$$H_q(p) = \frac{1}{q-1} \left(1 - \sum_s p_s^q \right) \quad (1)$$

Many known biodiversity measures, including the *Richness*, *Shannon* and *Simpson indices*, are obtained by the *deformed exponential transformation* of the Tsallis entropy with known values of q : $D_q(p) = e_q(H_q(p))$. When $x < \frac{1}{q-1}$ holds, the deformed exponential transformation of order q is defined as $e_q(x) = [1 + (1-q)x]^{\frac{1}{1-q}}$ that converges to the standard exponential when $q \rightarrow 1$.

The order q of these indeces can be seen as a penalizing parameter: a large value of q implies a reduction of the total diversity, given that rare species have less importance in determining the structure of the overall community.

Figure 2 shows the time series of the α -, β - and γ -diversity for $q = 0, 1, 2$. There is a decreasing trend in the α -diversity for $q = 0$ (*Richness*), while there is an increasing one for the β -diversity; these effects balance out and, as a result, the γ -

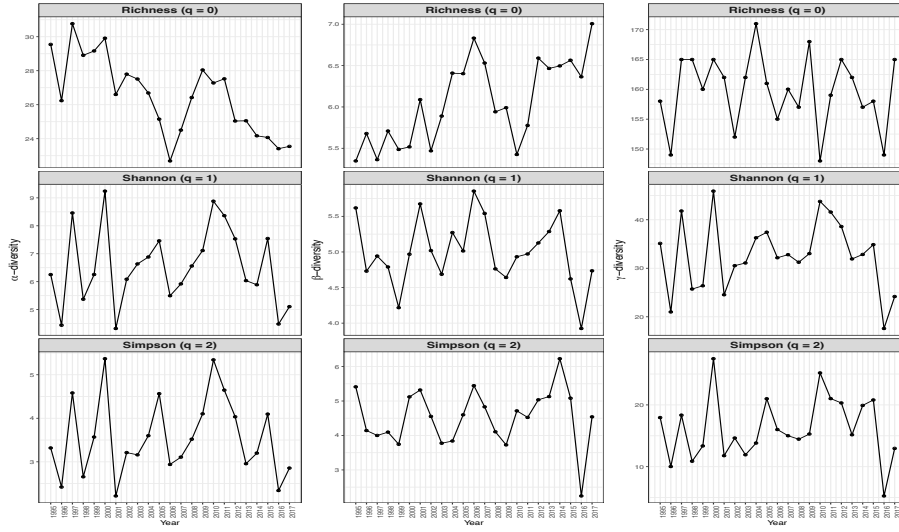


Fig. 2 α -, β - and γ -diversity time series for $q = 0, 1, 2$.

diversity is stable between 155 and 160 species across time. When penalized for the rare species (*Shannon* and *Simpson*), these trends in the α - and β -diversity are less evident and thus require further investigation.

3 Condecomp/Parafac

In order to reduce the complexity of the dataset we run the Condecomp/Parafac (CP) procedure³ [2, 6]. This method is very useful to analyze data which can be stored in a 3-way tensor lying in $\mathbb{R}^{I \times J \times K}$. It can be seen as a generalization of the simple PCA, where a new set of loadings c_{ks} are included to link the occasions to the components. In scalar form, the CP model can then be therefore formulated as:

$$x_{ijk} = \sum_{s=1}^S a_{is} b_{js} c_{ks} + e_{ijk}$$

for $i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$. and where e_{ijk} is the generic error term. We run the Parafac on the square root of the relative abundance index, excluding the species which are recorded less than the 5% of the total sampling occasions (40×23). Results showed that 4 components are able to explain $\simeq 49\%$ of the total variance. In particular, Table 1 shows that the first component seems to be related to the depth gradient along the continental slope with species like *M. merluccius*, *G. leioglossus*, *L. caudatus*, *I. coindetii* typically associated with the shelf-break habitat

³ implemented in the *ThreeWay* R-package [4].

and *A. foliacea*, *E. spinax* belonging to the middle-slope assemblage. The second component is instead associated with typical upper-slope species such as *G. melastomus*, *P. blennoides* and *N. norvegicus*. The third component can be interpreted as the behaviour pelagic (*S. pilchardius* and *E. encrasicolous*) vs benthic (*M. scolopax* and *G. leioglossus*). Coastal species such as *M. barbatus*, *O. vulgaris*, *P. erythrinus* are associated with the fourth component. Hence, shelf and edge stations are related to the first component and are more homogeneous between them than the ones belonging to the slope. The species assemblages structure do not seem to show main temporal changes during the study period.

Species	Comp.1	Comp.2	Comp.3	Comp.4
<i>M. merluccius</i>	-0.573			
<i>G. leioglossus</i>	-0.297		0.366	
<i>L. caudatus</i>	-0.219			
<i>I. coindetii</i>	-0.215			
<i>A. foliacea</i>	0.0196			
<i>E. spinax</i>	0.0183			
<i>G. melastomus</i>		-1.54		
<i>P. blennoides</i>		-0.803		
<i>N. norvegicus</i>		-0.557		
<i>E. encrasicolous</i>			-1.74	
<i>S. pilchardius</i>			-0.645	-0.846
<i>M. scolopax</i>			0.143	
<i>M. barbatus</i>				-0.904
<i>O. vulgaris</i>				-0.529
<i>P. erythrinus</i>				-0.484

Table 1 Loadings representing the association of the species to each of the four components.

4 Concluding remarks and further developments

In this work, we showed that the diversity of the demersal fauna living in front of the central Tyrrhenian Sea (Lazio coast) did not change significantly across time in terms of γ -diversity, although local changes (α -diversity) seem to be occurred and will be further investigated. The Parafac procedure showed that the components are related to the species occurring at different distance from the coast and to the species living in pelagic or benthic habits.

In future developments, the goal is to model the diversity indeces separately, also including chemical and physical factors such as nitrate, phosphate, and oxygen concentration, in addition to water salinity and temperature. We intend to provide possible extensions to [10, 1].

References

- [1] N. Abdalla, S. Banerjee, G. Ramachandran, M. Stenzel, and P. A. Stewart. Coastline kriging: A bayesian approach. *Annals of work exposures and health*, 62(7):818–827, 2018.
- [2] J. D. Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of “eckart-young” decomposition. *Psychometrika*, 35(3):283–319, 1970.
- [3] A. E. Gelfand, J. A. Silander, S. Wu, A. Latimer, P. O. Lewis, A. G. Rebelo, M. Holder, et al. Explaining species distribution patterns through hierarchical modeling. *Bayesian Analysis*, 1(1):41–92, 2006.
- [4] P. Giordani, H. A. Kiers, M. A. Del Ferraro, et al. Three-way component analysis using the r package threeway. *Journal of Statistical Software*, 57(7):1–23, 2014.
- [5] A. Guisan and W. Thuiller. Predicting species distribution: offering more than simple habitat models. *Ecology letters*, 8(9):993–1009, 2005.
- [6] R. A. Harshman. Foundations of the parafac procedure: Models and conditions for an” explanatory” multimodal factor analysis. 1970.
- [7] L. R. Iverson, A. M. Prasad, S. N. Matthews, and M. Peters. Estimating potential habitat for 134 eastern us tree species under six climate scenarios. *Forest Ecology and Management*, 254(3):390–406, 2008.
- [8] E. Marcon, I. Scotti, B. Hérault, V. Rossi, and G. Lang. Generalization of the partitioning of shannon diversity. *PloS one*, 9(3):e90289, 2014.
- [9] G. Midgley, L. Hannah, D. Millar, M. Rutherford, and L. Powrie. Assessing the vulnerability of species richness to anthropogenic climate change in a biodiversity hotspot. *Global Ecology and Biogeography*, 11(6):445–451, 2002.
- [10] S. Shirota, A. E. Gelfand, and S. Banerjee. Spatial joint species distribution modeling using dirichlet processes. *arXiv preprint arXiv:1711.05646*, 2017.
- [11] W. Thuiller, S. Lavergne, C. Roquet, I. Boulangeat, B. Lafourcade, and M. B. Araujo. Consequences of climate change on the tree of life in europe. *Nature*, 470(7335):531, 2011.
- [12] R. H. Whittaker. Vegetation of the siskiyou mountains, oregon and california. *Ecological monographs*, 30(3):279–338, 1960.

Latent Markov models with discrete separate cluster random effects on initial and transition probabilities

Modelli Latent Markov ad effetti casuali discreti e separati per le probabilità iniziali e di transizione

Giorgio E. Montanari and Marco Doretti

Abstract In this paper, a multilevel latent Markov model is proposed where through discrete random effects a two-way classification of sample unit clusters is obtained. Such random effects are defined as two discrete random variables, one of which affecting the initial probabilities and the other one affecting the transition probabilities of the Markovian latent process. We apply the model to data on health status of elderly patients clustered in nursing homes. The two-dimensional criterion used for classifying nursing homes highlights a plausible pattern, which can be useful for their management.

Abstract In questo articolo si propone un modello latent Markov multilivello in cui tramite effetti casuali discreti è possibile ottenere due raggruppamenti distinti di unità di secondo livello omogenee rispetto alle probabilità iniziali, da una parte, e alle probabilità di transizione del processo latente Markoviano, dall'altra. Il modello è applicato a dati sullo stato di salute di pazienti anziani raggruppati in case di cura. Il criterio bidimensionale utilizzato per raggruppare le case di cura fa emergere un possibile schema interpretativo, utile per la loro gestione.

Key words: multilevel model, clustered data, nursing homes.

Montanari Giorgio E.

Department of Political sciences, University of Perugia, Via A. Pascoli 20, 06123 Perugia, Italy
e-mail: giorgio.montanari@unipg.it

Doretti Marco

Department of Political sciences, University of Perugia, Via A. Pascoli 20, 06123 Perugia, Italy
e-mail: marco.doretti@unipg.it

1 Introduction

Latent Markov models are a well-established tool for the analysis of longitudinal categorical data. In particular, these models are tailored to settings where some repeatedly measured categorical variables are likely to be simultaneously influenced by an underlying latent trait of interest, which is assumed to be also a categorical variable and to probabilistically evolve over time like a first-order Markov chain [7, 1]. Due to the Markovianity assumption, the relevant parameters of such a latent process are its initial and transition probabilities.

In the last decades, numerous extensions of the basic latent Markov model have been proposed in the literature to incorporate additional information (like individual covariates) as well as other statistical model features specifically required by the application field considered. Among these there is multilevel modeling, which is in order when sample units are grouped within a number of different clusters. Examples of application of the multilevel latent Markov (MLM) model concern longitudinal datasets of patients grouped in nursing homes [2] and students in classes [5].

In MLM models, the multilevel structure of the data is quite often accounted for by means of discrete cluster random effects with a finite number of support points. In practice, this scheme creates groups every second level unit (i.e., cluster) belongs to, possibly in a dynamic fashion [3], with a certain probability vector. In this approach the support points characterizing each group of clusters collect cluster effects in a way such that effects on initial probabilities are univocally associated to effects on transition probabilities. This might be not entirely appropriate when effects on initial and transition probabilities have a different interpretation and a two-way classification of clusters - one according to the initial probabilities and one to the transition probabilities - is of interest.

Recently, a model with two separate cluster effects on initial and transition probabilities has been proposed in the literature which relies on continuous rather than discrete random effects [6]. Specifically, they are assumed to follow a bivariate normal distribution. Although it is not necessary to assume these two components are independent in the first place, clusters can in principle be compared, grouped or ranked along any of the two dimensions. In this paper, we extend the use of separate cluster effects to the class of discrete random effect models. This might be useful in those settings where the normality assumption, typically required by continuous effect models, is not appropriate.

The proposed model is applied to a longitudinal dataset collecting information on the health status of patients clustered in a number of nursing homes in Umbria (Italy). For this application, the above mentioned MLM model with separate effects is sensible, since the nature of nursing home effects on initial and transition probabilities is rather different. Specifically, the former reflect admission policies, whereas the latter are associated to the quality of the health care service provided; see [6] for a related discussion. Therefore, the two-dimensional cluster grouping procedure implied by the proposed approach seems in order.

2 MLM models with separate cluster random effects

Let $\mathbf{Y}_{hi}^{(t)} = (Y_{hi1}^{(t)}, \dots, Y_{hiJ}^{(t)})'$ denote the vector containing the J categorical response variables of the i -th unit of the h -th cluster at time occasion t . Every item can have a different number of response categories labelled from 1 to c_j , with $j = 1, \dots, J$. Each of the H clusters has its own number of units n_h , so that the overall sample size is $n = \sum_{h=1}^H n_h$. The number of measurement occasions $T_{hi} \leq T$ is unit-specific, with T denoting the maximum number of occasions. Response vectors can be collected across time, i.e. $\mathbf{Y}_{hi} = (\mathbf{Y}_{hi}^{(1)}, \dots, \mathbf{Y}_{hi}^{(T_{hi})})'$, and across cluster units, i.e. $\mathbf{Y}_h = (\mathbf{Y}_{h1}, \dots, \mathbf{Y}_{hn_h})'$. The same notation applies to vectors of individual covariates, that we denote by $\mathbf{X}_{hi}^{(t)}$. For each unit, the Markovian latent process $\mathbf{V}_{hi} = (V_{hi}^{(1)}, \dots, V_{hi}^{(T_{hi})})'$ is a collection of T_{hi} categorical unobserved variables with k_v latent states labelled from 1 to k_v . Cluster random effects on such a Markovian process are assumed to be time-invariant and are denoted by the vector $\mathbf{Z}_h = (U_h, W_h)'$, where U_h affects the initial probabilities and W_h affects the transition probabilities. Formally, U_h and W_h are two random variables taking k_u and k_w possible values. These values are denoted by ψ_u ($u = 1, \dots, k_u$) and ξ_w ($w = 1, \dots, k_w$), respectively. Therefore, \mathbf{Z}_h has $k_z = k_u k_w$ distinct support points \mathbf{z}_z ($z = 1, \dots, k_z$) arranged in the lexicographical order $\mathbf{z}_1 = (\psi_1, \xi_1)'$, $\mathbf{z}_2 = (\psi_1, \xi_2)'$, \dots , $\mathbf{z}_{k_z} = (\psi_{k_u}, \xi_{k_w})'$.

As typical in settings where the primary interest lies in modeling the latent trait, we assume covariates and cluster random effects to influence the individual Markovian processes \mathbf{V}_{hi} but not the measurement model, that is, the model for the response variables given the latent trait. Specifically, such a dependence structure leads to define the initial probabilities

$$\pi_{hi}(v|u) = P(V_{hi}^{(1)} = v | \mathbf{X}_{hi}^{(1)} = \mathbf{x}_{hi}^{(1)}, U_h = \psi_u),$$

the first-order transition probabilities

$$\pi_{hi}^{(t)}(v|\bar{v}, w) = P(V_{hi}^{(t)} = v | V_{hi}^{(t-1)} = \bar{v}, \mathbf{X}_{hi}^{(t)} = \mathbf{x}_{hi}^{(t)}, W_h = \xi_w)$$

($t = 2, \dots, T_{hi}$), and the conditional response probabilities

$$\phi_{jy_t v_t} = P(Y_{hij}^{(t)} = y_t | V_{hi}^{(t)} = v_t)$$

($j = 1, \dots, J$, $y_t = 1, \dots, c_j$, $v_t = 1, \dots, k_v$). Notice that conditional response probabilities are also time-invariant.

Here, we assume the latent trait $V_{hi}^{(t)}$ is unidimensional, with states corresponding to increasing intensities of a certain attribute (in our application, patient physical limitations). Due to the specifics of our application (see Section 3), we also assume that the greatest intensity state k_v is an absorbing one, meaning that units cannot visit other states after they visited it. Further, units cannot be in this state at the initial time. Clearly, this imposes the constraints $\pi_{hi}(k_v|u) = 0$, as well as $\pi_{hi}^{(t)}(v|k_v, w) = 0$ and $\pi_{hi}^{(t)}(k_v|k_v, w) = 1$, for every value of u and w and for $v = 1, \dots, k_v - 1$.

Because of the ordinal nature of the latent trait, for the initial probabilities a global logit parametrization of the type

$$\log \frac{\pi_{hi}(v+1|u) + \dots + \pi_{hi}(k_v-1|u)}{\pi_{hi}(1|u) + \dots + \pi_{hi}(v|u)} = \beta_{0v} + \psi_u + \mathbf{x}_{hi}^{(1)} \boldsymbol{\beta}_1 \quad (1)$$

($u = 1, \dots, k_u; v = 1, \dots, k_v - 2$) is used. The parameter β_{0v} is an intercept varying with the logit equations, ψ_u represents the effect due to cluster h belonging to latent group u , while $\boldsymbol{\beta}_1$ is a constant vector of regression coefficients related to individual covariates. The global logit parametrization requires the sequence $\beta_{01}, \dots, \beta_{0k_v-2}$ be non-increasing to ensure the cumulative probabilities are non-decreasing. Further, to identify the model $\psi_1 = 0$ is set.

With regard to the transition probabilities, a rectangular $(k_v - 1 \times k_v)$ matrix has to be parametrized since the last row is constrained as illustrated above. For these probabilities, we here specify a global logit model similar to the previous one, that is,

$$\log \frac{\pi_{hi}^{(t)}(v+1|\bar{v}, w) + \dots + \pi_{hi}^{(t)}(k_v|\bar{v}, w)}{\pi_{hi}^{(t)}(1|\bar{v}, w) + \dots + \pi_{hi}^{(t)}(v|\bar{v}, w)} = \gamma_{0\bar{v}} + \gamma_{1v} + \xi_w + \mathbf{x}_{hi}^{(t)} \boldsymbol{\gamma}_2 \quad (2)$$

with $w = 1, \dots, k_w$, $\bar{v} = 1, \dots, k_v - 1$, $v = 1, \dots, k_v - 1$ and $t = 2, \dots, T_{hi}$. Again, the sequence $\gamma_{11}, \dots, \gamma_{1k_v-1}$ must be non-increasing, whereas $\boldsymbol{\gamma}_2$ is a vector of time-invariant regression parameters - constant across logit equations - for the effect of individual-level covariates. The ξ_w parameters are random intercepts representing cluster group effects on transition probabilities. These effects are also time-invariant and constant across logit equations. The constraints $\gamma_{01} = 0$ and $\xi_1 = 0$ are imposed for identifiability.

Finally, setting $\tilde{\pi}_z = P(\mathbf{Z}_h = \mathbf{z})$ ($z = 1, \dots, k_z$), a multinomial logit parametrization without covariates is adopted for the joint distribution of the random effects, i.e.

$$\log \frac{\tilde{\pi}_z}{\tilde{\pi}_1} = \rho_z \quad z = 2, \dots, k_z.$$

To reduce the overall complexity, maximum likelihood estimation is performed via a two-step algorithm; see [4]. In the first step, the time-invariant conditional response probabilities (i.e., the measurement model) are estimated by a latent class model pooling all the individual records together. In the second step, the estimated conditional response probabilities are plugged in the likelihood function of the data and a direct maximization of it returns the estimates of the remaining model parameters.

3 Model results for nursing home grouping

As stated in the Introduction, the proposed MLM model with separate effects is applied to multilevel longitudinal data concerning the physical limitations of elderly

patients (hereafter disability) hosted in Umbrian nursing homes. The dataset considered for this application refers to the years 2015-2016 and is formed by $N = 4,862$ observations relative to $n = 2,039$ individuals grouped in $H = 49$ nursing homes.

With regard to the response variables, $J = 10$ ordinal items are considered which measure patient difficulties in taking common actions like walking, getting dressed or maintaining personal hygiene. All the items have $c_j = c = 6$ categories. In this setting, the Markovian unobserved process under investigation represents the overall level of disability, justifying the assumption that the latent trait is one-dimensional and ordinal (see Section 2) as well as the global logit parametrization in Equations (1) and (2). In both these equations, patient age (in years) and gender (a binary variable taking value 1 for males) are included as individual covariates.

In the examined dataset, patients are followed up at most for $T = 3$ measurements taken approximately every six months. However, a number of patients have data only for one or two occasions due to dropout, which might be caused by death or other reasons. While the latter can be reasonably assumed to be unrelated to patient health status, death is clearly the worse physical health condition. As such, it can be considered as an outcome of interest rather than a cause of missing data. In particular, in the proposed model death is univocally associated to the k_v -th latent state, which, as mentioned in Section 2, is an absorbing state patients can not be at at the first occasion. To embed information on patient death in the model, a data expansion is performed where the 525 patients dropping out due to death are assigned an extra observation with all items set to the additional response category $c + 1 = 7$, that represents the most extreme level of difficulty in every-day activities. From a conceptual standpoint, this corresponds to $\phi_{j,c+1,k_v} = 1$ ($j = 1, \dots, J$).

In latent variable models, model selection is concerned with the choice of the number of latent states. Here, a three-fold criterion is needed since latent groups are created at the individual as well as at the cluster level, and the triplet (k_v, k_u, k_w) has to be specified. In this regard, $k_v = 6$ was selected in the first estimation step, and, as an illustration, we here report results for the pair $k_u = 2$ and $k_w = 2$.

Results for the measurement model and for the effects of individual covariates are in line with those obtained in similar settings; see for example [6]. With regard to the measurement model, the $k_v = 6$ identified latent states are naturally ordered according to the level of seriousness of patient disability, with the last one corresponding to the patient death (see Section 3). Note that these results were obtained without explicitly imposing any order constraint or the constraints $\hat{\phi}_{j,c+1,k_v} = 1$ ($j = 1, \dots, J$). As for individual covariates, we obtain $\hat{\beta}_1 = (0.019, -0.384)'$ and $\hat{\gamma}_2 = (0.021, 0.150)'$. As expected, the two coefficients of age are positive, meaning that older patients are more likely to be in worse disability conditions at the first occasion as well as to move towards more serious states at the following ones. In addition, coefficients of gender indicate that males typically have a better health status at baseline but tend to migrate to worse conditions more likely than females. All these coefficients, which are stable across model combinations, are significant at the 5% level.

The parameter estimates of nursing home group effects are $\hat{\psi}_2 = 1.038$ and $\hat{\xi}_2 = -0.222$, while the estimated joint distribution of $(U_h, W_h)'$, arranged in matrix form,

is

$$\begin{pmatrix} \hat{\pi}_1 & \hat{\pi}_2 \\ \hat{\pi}_3 & \hat{\pi}_4 \end{pmatrix} = \begin{pmatrix} 0.617 & 0.120 \\ 0.263 & 0.000 \end{pmatrix}.$$

The value of $\hat{\psi}_2$ shows that the second of the two groups formed with respect to initial probabilities include nursing homes with a higher propensity to host patients in worse conditions - age and gender being equal. Conversely, $\hat{\xi}_2 < 0$ implies that the second of the two groups formed with respect to transition probabilities corresponds to nursing homes with a greater ability of preserving their patients from a disability worsening. However, this division according to effects on transition probabilities seems to take place only for the nursing homes in the first initial probability group. A possible explanation of these results is that nursing homes in the second initial probability group house patients with a higher level of disability in the first occasion and no distinction among them is possible with regards to the ability to avoid disability worsening. Although rather plausible overall, these preliminary results might be enhanced by further analyses and refinements both in the estimation and in the model selection phase.

Acknowledgements This research is funded by a grant of Fondazione Cassa di Risparmio di Perugia and University of Perugia.

References

- [1] F. Bartolucci, A. Farcomeni, and F. Pennoni. *Latent Markov Models for Longitudinal Data*. Statistics in the Social and Behavioural Sciences. Chapman & Hall/CRC, 2013.
- [2] F. Bartolucci and M. Lupporelli. The multilevel latent Markov model. In *Proceedings of the International Workshop on Statistical Modelling, Barcelona*, pages 93–98, 2007.
- [3] F. Bartolucci and M. Lupporelli. Pairwise likelihood inference for nested hidden Markov chain models for multilevel longitudinal data. *Journal of the American Statistical Association*, 111:216–228, 2016.
- [4] F. Bartolucci, G. E. Montanari, and S. Pandolfi. A comparison of some estimation methods for latent Markov models with covariates. In *Proceedings of COMPSTAT 2014 - 21st international conference on computational statistics*, pages 531–538, 2014.
- [5] F. Bartolucci, F. Pennoni, and G. Vittadini. Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics*, 36:491–522, 2011.
- [6] G. E. Montanari, M. Doretti, and F. Bartolucci. A multilevel latent Markov model for the evaluation of nursing homes' performance. *Biometrical Journal*, 60(5):962–978, 2018.
- [7] L. M. Wiggins. *Panel analysis: Latent probability models for attitude and behavior processes*. Jossey-Bass, 1973.

Unsuitability of likelihood-based asymptotic confidence intervals for Response-Adaptive designs in normal homoscedastic trials

Inadeguatezza degli intervalli di confidenza asintotici basati sulla verosimiglianza per disegni Response-Adaptive in caso di risposte normali omoschedastiche

Marco Novelli and Maroussa Zagoraiou

Abstract The present paper discusses the drawbacks of likelihood-based asymptotic confidence intervals in sequential clinical trials for treatment comparisons managed via Response-Adaptive randomization. By pushing forward the results obtained in [2], where the authors have shown the weakness of this approach in the hypothesis testing framework for normal homoscedastic trials, in this paper we provide the conditions under which the asymptotic confidence intervals degenerate, also showing via examples and simulations how the choice of the target allocation may severely undermine the inferential precision of the classical procedure.

Abstract *In questo lavoro viene discussa l'inadeguatezza degli intervalli di confidenza asintotici basati sulla verosimiglianza nell'ambito delle prove cliniche sequenziali di tipo Response-Adaptive. Partendo dai risultati ottenuti in [2], dove sono state evidenziate le limitazioni della statistica di Wald nell'ambito dei test asintotici, vengono qui illustrate le criticità di un approccio inferenziale basato sulla verosimiglianza in termini di intervalli di confidenza asintotici con risposte normali omoschedastiche. In particolare sono discusse, anche attraverso esempi e simulazioni, le condizioni sotto cui tali intervalli possono degenerare, facendone quindi fallire l'applicabilità.*

Key words: Asymptotic normality, confidence intervals, sequential clinical trials, target allocations.

Department of Statistical Sciences, University of Bologna
Via Belle Arti 41, Bologna e-mail: marco.novelli4@unibo.it

Department of Statistical Sciences, University of Bologna
Via Belle Arti 41, Bologna e-mail: maroussa.zagoraiou@unibo.it

1 Introduction

Adaptive experiments are sequential procedures in which the decision about how to proceed is made according to a pre-established rule that makes use of the information accrued along the way. They are nowadays considered as a *gold standard* in the clinical context, especially in phase-III trials for treatment comparisons, where the ethical goal of maximizing the subjects care often conflicts with the statistical aim of drawing correct inferential conclusions. To overcome this trade-off, many authors suggested target allocations of the treatments representing a valid compromise among ethics and inference (see e.g. [3, 6]). Generally, these targets depend on the unknown model parameters and they can be approached asymptotically by using suitable Response-Adaptive (RA) randomization procedures, namely sequential allocation rules in which the treatment allocation probabilities change on the basis of earlier responses and past assignments in order to progressively approach the chosen target. A well-known example is the efficient randomized-adaptive design (ERADE) suggested in [5].

Even if RA procedures induce a complex dependence structure between the outcomes, several authors provided the conditions under which the classical likelihood-based asymptotic inference is still valid. However, in many circumstances such approach may present critical drawbacks, especially when the chosen target is characterized by a strong ethical component, as showed in the hypothesis testing framework (see [2, 4]). This paper explores the inadequacy of the classical likelihood-based approach in deriving reliable confidence intervals by stressing the crucial role played by the target that could i) compromise the quality of the CLT approximation of the standard MLEs and ii) lead to a vanishing Fisher information, thus severely undermining the inferential precision.

2 Likelihood-based asymptotic inference

Consider a clinical trial where patients come sequentially and are assigned to one of two competing treatments, say A and B . At each step i , let $\delta_i = 1$ if the i th patient is allocated to A and 0 otherwise. Let also Y_i be the normally distributed outcome of the corresponding subject, where $E(Y_i | \delta_i = 1) = \theta_A$ and $E(Y_i | \delta_i = 0) = \theta_B$ denote the treatment effects and $V(Y_i) = \sigma^2$ is the unknown common variance. After n patients, let $N_{An} = \sum_{i=1}^n \delta_i$ and $N_{Bn} = n - N_{An}$ be the assignments to the treatments, so that $\pi_n = n^{-1}N_{An}$ is the allocation proportion to A . Then, the MLEs of the treatment effects coincide with the sample means, that is $\hat{\theta}_{An} = N_{An}^{-1} \sum_{i=1}^n \delta_i Y_i$ and $\hat{\theta}_{Bn} = N_{Bn}^{-1} \sum_{i=1}^n (1 - \delta_i) Y_i$, while $M_n = \sigma^{-2} \text{diag}(\pi_n; [1 - \pi_n])$ is the normalized Fisher information matrix associated to the parameters of interest.

The inferential objective typically consists in estimating the superiority of a new treatment wrt the gold standard (say A wrt B), namely the difference between the treatment effects $\Delta = \theta_A - \theta_B \in \mathbb{R}$. Without loss of generality, in what follows we

assume “the-larger-the-better” scenario, that is higher responses are assumed to be preferable for the patient.

The trade-off between ethical concerns and inferential precision can be managed through a suitable RA randomization procedure that skews the assignment to the best treatment and converges to a properly chosen target allocation to A , say $\rho = \rho(\Delta)$. For example, in the context of normal homoscedastic outcomes, Atkinson and Biswas [1] suggested $\rho_N(\Delta) = \Phi(\Delta/T)$, where Φ is the cdf of the standard normal and $T > 0$ is a tuning parameter managing the ethical skew (i.e., for low values of T the target strongly skews the assignments to the better treatment, while as T grows ρ tends to be balanced).

In general, the target should satisfy (for any $T > 0$) the following assumptions:

- A1: $\rho : \mathbb{R} \rightarrow (0, 1)$ is a continuous symmetric function such that $\rho(-\Delta) = 1 - \rho(\Delta)$, which guarantees that both treatments are considered similarly;
- A2: ρ is increasing in Δ , meaning that the best treatment should be favored increasingly as its superiority grows.

Given a chosen target ρ , RA randomized procedures can be adopted to converge to it. These are sequential procedures such that at each step n the treatment effects $\theta = (\theta_A; \theta_B)^t$ are estimated by means of $\hat{\theta}_n = (\hat{\theta}_{An}; \hat{\theta}_{Bn})^t$ and the target is estimated accordingly, i.e., $\hat{\rho}_n = \rho(\hat{\Delta}_n)$, where $\hat{\Delta}_n = \hat{\theta}_{An} - \hat{\theta}_{Bn}$. Thus, the allocation of the next patient is forced to the better treatment and the allocation proportion π_n progressively approaches ρ as n grows.

Due to the complex dependence structure induced by RA procedures, the distribution of the MLEs is not the same as in the non-sequential setting; however, given a target satisfying assumptions A1-A2, if the RA design is chosen such that $\lim_{n \rightarrow \infty} \pi_n = \rho(\Delta)$ a.s., then $\lim_{n \rightarrow \infty} M_n = M = \sigma^{-2} \text{diag}(\rho(\Delta); [1 - \rho(\Delta)])$ a.s. and the consistency of the MLEs is guaranteed along with their asymptotic normality. This allows for the standard asymptotic inference based on the likelihood, since

$$\sqrt{n}(\hat{\Delta}_n - \Delta) \hookrightarrow \mathcal{N}\left(0, \frac{\sigma^2}{\rho(\Delta)[1 - \rho(\Delta)]}\right), \quad (1)$$

so that the $(1 - \alpha)\%$ asymptotic confidence interval for Δ is

$$CI(\Delta)_{1-\alpha} = \left(\hat{\Delta}_n \pm \frac{z_{1-\alpha/2}}{\sqrt{n}} \left[\frac{\hat{\sigma}_n^2}{\rho(\hat{\Delta}_n)[1 - \rho(\hat{\Delta}_n)]} \right]^{\frac{1}{2}} \right), \quad (2)$$

where z_α is the α -percentile of Φ and $\hat{\sigma}_n^2$ is a consistent estimator of σ^2 .

3 Inadequacy of asymptotic confidence intervals for RA designs

In this section we discuss the inadequacy of the likelihood-based asymptotic inference in deriving reliable confidence intervals under RA randomization, showing

how the chosen target plays a central role in terms of inferential precision. From (1), it is clear that, if the target has a high ethical component the asymptotic variance of $\hat{\Delta}_n$ tends to diverge and the quality of the CLT approximation could be worsened, thus compromising any likelihood-based inferential procedure. Indeed, the width of $CI(\Delta)_{1-\alpha}$ in (2) is proportional to

$$\frac{\hat{\sigma}_n^2}{\rho(\hat{\Delta}_n)[1 - \rho(\hat{\Delta}_n)]}$$

so that the interval could degenerate when the chosen target tends to either 0 or 1 as the difference between the treatment effects grows (i.e., as Δ diverges). This anomalous behaviour may be exacerbated by the effect of the tuning parameter, since the quality of the CLT approximation as well as the divergence of the asymptotic variance in (1) are strongly related to the choice of T . For example, adopting ρ_N , T induces a re-scaling effect by shrinking (for $T > 1$) or expanding (for $T < 1$) the treatment difference Δ . Hence small values of T may damage the CLT approximation even for small values of Δ .

As we will show in the next Section, the above-mentioned behaviour can be partially circumvented by re-scaling the target. Indeed, by letting $\tilde{\rho}(x) = 1 - b + \rho(x)(2b - 1)$ with $b \in (1/2; 1)$, then $\lim_{x \rightarrow \infty} \tilde{\rho}(x) = b < 1$ and, at the same time, $\lim_{x \rightarrow -\infty} \tilde{\rho}(x) = 1 - b > 0$, thus avoiding degenerate scenarios.

3.1 Simulation study

In order to highlight how the choice of the target allocation and the tuning parameter T could severely undermine the precision of the likelihood-based inferential procedures, we perform a simulation study with 500000 simulated trials in the case of homoscedastic normal responses with $\sigma^2 = 1$ and $n = 250$, where ERADE is employed. We compare the results adopting ρ_N and its re-scaled version $\tilde{\rho}_N$ (with $b = 0.95$) for several values of T . Figure 1 shows the simulated distributions of the MLE $\hat{\Delta}_n$ as Δ and T vary, while Table 1 and Table 2 summarize the behaviour of the simulated 95% asymptotic confidence intervals for Δ : Lower (L) and Upper (U) bounds are computed by averaging the endpoints of the simulated trials, while the analytical values derived by (2) are reported within brackets.

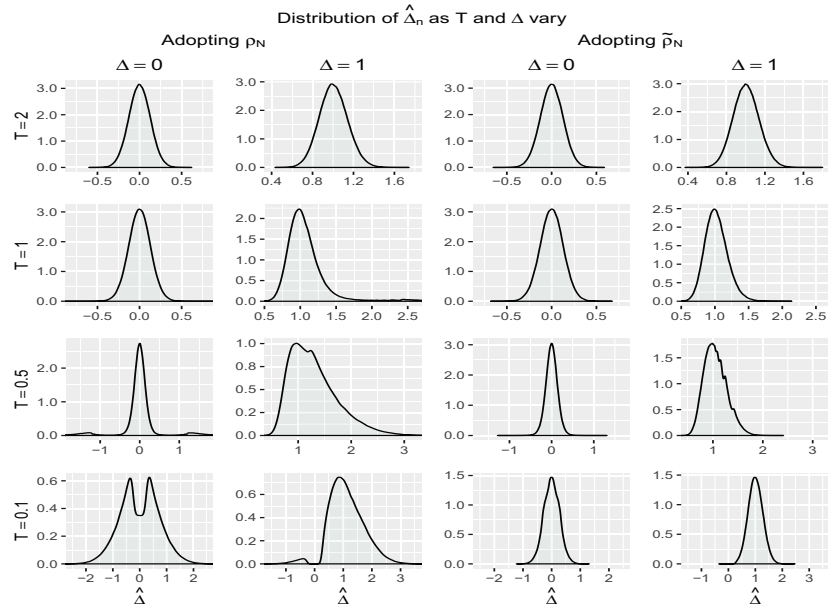


Fig. 1 Simulated distribution of $\hat{\Delta}_n$ adopting ρ_N and $\tilde{\rho}_N$ as T and Δ vary.

Taking into account ρ_N , low values of T seriously affect the CLT approximation, leading to an evident increase of the density on the tails, whereas the adoption of $\tilde{\rho}_N$ considerably reduces this effect. For $\Delta = 1$, the distribution of the MLE tends to present a positive skewness regardless of T , with stronger departure from normality for ρ_N combined with smaller values of the tuning parameter. Again, this effect is partially soothed by the adoption of the re-scaled target $\tilde{\rho}_N$.

Table 1 The behaviour of simulated asymptotic $CI(\Delta)_{0.95}$ adopting ρ_N as T and Δ vary.

		Δ							
		0		1		2		3	
T		L	U	L	U	L	U	L	U
		(-0.25)	(0.25)	(0.73)	(1.27)	(1.66)	(2.34)	(2.50)	(3.50)
2		-0.25	0.25	0.71	1.30	1.59	2.44	2.41	3.73
1		-0.25	0.25	0.66	1.50	0.52	3.98	-7.19	13.44
		(-0.25)	(0.25)	(0.66)	(1.34)	(1.17)	(2.83)	(-0.38)	(6.38)
0.5		-1.44	1.44	-209	211	-25498	25502	-443988	443994
		(-0.25)	(0.25)	(0.17)	(1.83)	(-20.03)	(24.03)	(-3943.48)	(3949.48)
0.1		-709456	709456	$-\infty$	∞	$-\infty$	∞	$-\infty$	∞
		(-0.25)	(0.25)	(-4.5×10^{10})	(4.5×10^{10})	$(-\infty)$	(∞)	$(-\infty)$	(∞)

L, U: average lower and upper simulated bounds (analytical endpoints in brackets).

It is clear from Table 1 that, under ρ_N , even if the simulated confidence bounds are quite close to the analytical ones for small values of Δ and $T \geq 1$, the skewness of the MLE's distribution severely worsen the quality of the CLT approximation as

Δ increases. Moreover, small values of T strongly undermine the accuracy of the intervals, that tend to rapidly diverge even for small Δ .

Table 2 The behaviour of simulated asymptotic $CI(\Delta)_{0.95}$ adopting $\tilde{\rho}_N$ as T and Δ vary.

		Δ							
		0		1		2		3	
		L	U	L	U	L	U	L	U
T	2	−0.25	0.25	0.71	1.30	1.61	2.41	2.47	3.57
		(−0.25)	(0.25)	(0.74)	(1.26)	(1.69)	(2.31)	(2.60)	(3.40)
	1	−0.25	0.25	0.68	1.37	1.48	2.58	2.33	3.68
		(−0.25)	(0.25)	(0.69)	(1.31)	(1.52)	(2.48)	(2.44)	(3.56)
	0.5	−0.25	0.25	0.55	1.56	1.38	2.62	2.31	3.69
		(−0.25)	(0.25)	(0.52)	(1.48)	(1.43)	(2.57)	(2.43)	(3.57)
	0.1	−0.43	0.43	0.43	1.59	1.37	2.62	2.31	3.69
		(−0.25)	(0.25)	(0.43)	(1.57)	(1.43)	(2.57)	(2.43)	(3.57)

L, U: average lower and upper simulated bounds (analytical endpoints in brackets).

As it can be seen in Table 2, the adoption of the re-scaled target $\tilde{\rho}_N$ considerably improves the inferential precision and the reliability of the intervals. Indeed, in every scenario, the simulated and the analytical endpoints are almost coincident, with only small discrepancies. Clearly, the width of the intervals slightly increases as the ethical impact becomes more relevant (i.e., for very small values of T) but not at the expense of the inferential accuracy.

Acknowledgments Research supported by the Italian Ministry of Education, University and Research under PRIN 2015 Environmental processes and human activities: capturing their interactions via statistical methods (EphaStat)

References

1. Atkinson, A.C., Biswas, A. Bayesian adaptive biasedcoin designs for clinical trials with normal responses. *Biometrics*, **61**, 118-125 (2005)
2. Baldi Antognini, A., Vagheggini, A., Zagoraiou, M. Is the classical Wald test always suitable under response-adaptive randomization? *Stat. Meth. Med. Res.*, **27**, 2294-2311 (2016)
3. Baldi Antognini, A., Novelli, M., Zagoraiou, M. Optimal designs for testing hypothesis in multiarm clinical trials. *Stat. Meth. Med. Res.*, (2018) doi: 0962280218797960.
4. Baldi Antognini, A., Vagheggini, A., Zagoraiou, M., Novelli, M. A new design strategy for hypothesis testing under response adaptive randomization. *Electronic Journal of Statistics*, **12**, 2454-2481 (2018)
5. Hu, F., Zhang, L.X., He, X. Efficient randomized-adaptive designs. *Ann. Stat.*, **37**, 2543-2560 (2009)
6. Rosenberger, W.F., Stallard, N., Ivanova, A., Harper, C.N., Ricks, M.L. Optimal adaptive designs for binary response trials. *Biometrics*, **57**, 909-913 (2001)

Local Hypothesis Testing for Functional Data: Extending False Discovery Rate to the Functional Framework

*Verifica locale delle ipotesi nell'ambito dei dati
funzionali: estensione della nozione di False Discovery
Rate al contesto funzionale*

Niels Asken Lundtorp Olsen, Alessia Pini, and Simone Vantini

Abstract A topic which is becoming more and more popular in Functional Data Analysis is local inference, i.e., the continuous statistical testing of a null hypothesis along the domain. This can be seen as an extreme case of the multiple comparison problem. During the talk, we will define and discuss the notion of False Discovery Rate (FDR) in the setting of functional data. We will then introduce a new procedure (i.e., a continuous version of the Benjamini-Hochberg procedure) able to control the FDR over the functional domain, describe its properties in terms of control of the Type-I error probability and of consistency. The proposed method will be applied to satellite measurements of Earth temperature with the aim of identifying the regions of the planet where temperature has significantly increased in the last decades.

Abstract *Un ambito di ricerca divenuto sempre più oggetto di indagine nell'ambito dell'analisi di dati funzionali è l'inferenza locale ovvero la verifica delle ipotesi diffusa lungo il dominio che può essere interpretata come una condizione estrema di test multipli. Durante la relazione orale verrà proposta e approfondita la nozione di False Discovery Rate (FDR) nell'ambito di dati funzionali. Verrà anche introdotta una nuova procedura (una versione continua della procedura di Benjamini-Hochberg) in grado di garantire il controllo del FDR lungo il dominio e verranno poi descritte le sue proprietà in termini di controllo della probabilità di errore del primo tipo e di consistenza. La metodologia proposta verrà in fine applicata all'analisi di dati satellitari con lo scopo di identificare le regioni del pianeta nelle*

Niels Asken Lundtorp Olsen

Dept. of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark, e-mail: niels.olsen@math.ku.dk

Alessia Pini

Dept. of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milano, Italy, e-mail: alessia.pini@unicatt.it

Simone Vantini

MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, e-mail: simone.vantini@polimi.it

quali si è osservato un riscaldamento statisticamente significativo nell'arco degli ultimi decenni.

Key words: functional data, local inference, null hypothesis testing, false discovery rate, Benjamini Hochberg.

1 Motivation

Statisticians are more and more confronted with the analysis of *complex* data, where *complexity* often takes the form of a data analysis that pertains to analyzing data that are represented with abstract mathematical constructs, often belonging to some space on which a Hilbert structure is assumed (i.e., Object-oriented Data Analysis, [14]). For example, the continuous and outstanding advances of measurement technologies have enabled the collection and storage of high-resolution data which can often be modeled as smooth functions (e.g., curves or surfaces). This kind of data are at the basis of functional data analysis (FDA) which is a well-known lively and expanding research area of modern statistics. In FDA, the classical concept for scalar or multivariate random variable is indeed replaced by the concept of functional random variable. Consequently, in FDA the typical data set is not made of numbers or Euclidean vectors, but a collection of functions embedded in a suitable functional Hilbert space meant to formalize application-specific relations between sample units ([21, 9]). While FDA is expanding rapidly with recent applications of FDA techniques in different and many fields of science being countless, the theoretical study of statistical tools for making inference in such spaces is still a lively area of methodological investigation ([8, 23, 6, 22, 2, 11, 4, 5, 12, 3, 10, 18]).

2 Mathematical framework

A topic which is becoming more and more popular in Functional Data Analysis is local inference, i.e., the continuous statistical testing of a null hypothesis along the domain. As a toy example, think at the functional version of a standard two-sample test which can be formalized as follows. Let $\{\xi_{1i}\}_{i=1,\dots,n_1}$ and $\{\xi_{2i}\}_{i=1,\dots,n_2}$ be two independent random samples of functions in $L^2(T)$, where $i = 1, \dots, n_j$ is the unit index, $j = 1, 2$ is the population index, $T = (a, b) \subset \mathcal{R}$, and $L^2(T)$ is the Hilbert space of squared-integrable functions on the interval T . Let μ_1 and μ_2 be the functional means of the two populations. We want to perform the following null hypothesis significance test:

$$H_0 : \mu_1 = \mu_2 \text{ against } H_1 : \mu_1 \neq \mu_2.$$

The classical approach to face this problem is the identification of a *global* criterion to reject (or not to reject) the null hypothesis based on the definition of a suitable *p*-value. The goal of local inference is different: in the toy example above, it is indeed the identification of a *local* criterion to point-wise reject (or not to reject) the null hypothesis along the domain such to be able to select the portions of the domain where the two means are significantly different. To reach this goal, a continuous infinity of tests (in bijective correspondence with the points of the domain) would be defined, performed, and summarised in the so-called *p*-value function. The principal issue in this topic is the infinite amount of performed tests, which can be seen as an extreme case of the multiple comparison problem.

3 Methods

Recent literature in local inference ([1, 19, 20, 16, 21, 24, 25]) has faced this challenge by focusing on the definition of *p*-value functions (i.e., functions associating a *p*-value to each point of the domain) aiming at controlling the Family Wise Error Rate (FWER, i.e, the probability that the Lebesgue measure of the random region of the domain in which the null hypothesis is true but rejected is positive). During the talk, we will define a novel notion of False Discovery Rate (FDR) in the setting of functional data defined on a compact set. We will also introduce a continuous version of the Benjamini-Hochberg procedure able to map a generic unadjusted *p*-value function (i.e., a *p*-value which is point-wise exact) into an adjusted *p*-value function able to control the FDR over the entire functional domain. In detail, we will discuss some general theoretical conditions under which the adjusted *p*-value function produced by the novel procedure is consistent and also provided with the control of functional FDR. By means of simulations we will show that, at the cost of a weaker control of the probability of Type-I error, this new proposal generally presents remarkably larger power than other state-of-the-art procedures aiming at controlling the FWER.

Finally, to show the flexibility of the approach, the functional Benjamini-Hochberg procedure hereby presented will be applied to the analysis of satellite remote sensing measurements of Earth temperature. In detail, we will identify the regions of the planet where temperature has significantly increased in the period 1983-2007 while controlling the FDR over the surface of the planet. To this purpose, we will model yearly temperature maps as functional data defined on a spherical domain and possibly affected by a site-specific linear trend. Then, we will test the significance of a functional linear trend of a function-on-scalar regression model explaining yearly temperature maps as a function of time:

$$T_i(s) = \beta_0(s) + \beta_1(s)t_i + \varepsilon_i(s), \quad s \in S^2,$$

with

$$H_0 : \beta_1(s) = 0 \quad \forall s \in S^2 \quad \text{and} \quad H_1 : \exists s \in S^2 : \beta_1(s) > 0.$$

Theory, simulations, and details about the climate change application presented in the talk are fully detailed in [15].

References

1. Abramowicz, K., De Luna, S., Hger, C., Pini, A., Schelin, L., Vantini, S. (2018): Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament, *Scandinavian Journal of Statistics*, Vol. 45(4), pp. 1036-1061.
2. Cardot, H., Prchal, L., Sarda, P. (2007): No effect and lack-of-fit permutation tests for functional regression. *Computational Statistics*. **22**, 371–390.
3. Corain, L., Melas, V. B., Pepelyshev, A., Salmaso, L. (2014): New insights on permutation approach for hypothesis testing on functional data. *Advances in Data Analysis and Classification*. **8**, 339–356.
4. Cox, D. D., Lee, J. S. (2008): Pointwise testing with functional data using the Westfall–Young randomization method. *Biometrika*. **95**, 621–634.
5. Cuesta-Albertos, J. A., Febrero-Bande, M. (2010): A simple multiway ANOVA for functional data. *Test*. **19**, 537–557.
6. Cuevas, A., Febrero, M., Fraiman, R. (2004): An ANOVA test for functional data. *Computational statistics & data analysis*. **47**, 111–122.
7. Egozcue, J. J., Díaz-Barrero, J. L., Pawlowsky-Glahn, V. (2006): Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica*. **22**, 1175–1182.
8. Fan, J., Lin, S. K. (1998): Test of significance when data are curves. *Journal of the American Statistical Association*. **93**, 1007–1021.
9. Ferraty, F., Vieu, P. (2006): *Nonparametric functional data analysis: theory and practice*, Springer, New York.
10. Galeano, P., Esdras, J., Lillo, R.E. (2015): The Mahalanobis distance for functional data with applications to classification. *Technometrics*. **57**(2), 281–291.
11. Hall, P., Van Keilegom, I. (2007): Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*. **17**, 1511.
12. Horváth, L., Kokoszka, P. (2012): *Inference for functional data with applications*. Springer, New York.
13. Hsing, T., Eubank, R. (2015): *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.
14. Marron, J. S., Alonso, A. M. (2014): Overview of object oriented data analysis. *Biometrical Journal*.
15. Olsen, N., Pini, A., Vantini, S. (2019): A Benjamini-Hochberg procedure for local inference in functional data analysis, MOX Report, Dept of Mathematics, Politecnico di Milano.
16. Pini, A., Spreafico, L., Vantini, S., Vietti, A. (2019): Multi-aspect local inference for functional data: analysis of ultrasound tongue profiles. *Journal of Multivariate Analysis*, Vol. 170, pp. 162-185.
17. Pini, A., Stamm, A., Vantini, S. (2018): Hotelling's T^2 in separable Hilbert spaces, *Journal of Multivariate Analysis*, Vol. 167, pp. 284-305.
18. Pini, A., Vantini, S. (2016): The interval testing procedure: A general framework for inference in functional data analysis. *Biometrics*. **72**, 3, 835-845
19. Pini, A., Vantini, S. (2017): Interval-wise Testing for Functional Data, *Journal of Nonparametric Statistics*, Vol. 29(2), pp. 407-424.
20. Pini, A., Vantini, S., Colosimo, B. M., Grasso, M. (2017): Domain-Selective Functional ANOVA for Supervised Statistical Profile Monitoring of Signal Data, *Journal of the Royal Statistical Society ? Series C*, Vol. 67(1), pp. 55-81.
21. Ramsay, J. O., Silverman, B. W. (2005): *Functional data analysis*. Springer, New York.

22. Shen, Q., Faraway, J. (2004): An F test for linear models with functional responses. *Statistica Sinica*. 1239–1257.
23. Spitzner, D. J., Marron, J. S., Essick, G. K. (2003): Mixed-model functional ANOVA for studying human tactile perception. *Journal of the American Statistical Association*. **98**, 263–272.
24. Vsevolozhskaya, O. A., Greenwood, M. C., Bellante, G. J., Powell, S. L., Lawrence, R. L., Repasky, K. S. (2013). Combining functions and the closure principle for performing follow-up tests in functional analysis of variance. *Computational Statistics & Data Analysis*, Vol. 67, pp. 175?184.
25. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage*, Vol. 92, pp. 381?397.

Educational mismatch and attitudes towards migration in Europe

Disallineamento fra formazione e lavoro e atteggiamenti verso le migrazioni in Europa

Marco Guido Palladino¹ and Emiliano Sironi²

Abstract In this paper we focus on the relationship between educational mismatch and attitudes towards migration. Educational mismatch occurs when the required level of education for a particular job diverges from the employees' attained level of education. Overeducated individuals are those with an education higher than that needed for the job. Conversely, undereducated workers have an education level lower than required. Using data from the round 7 of the European Social Survey (ESS) this paper supports the hypothesis that overeducated individuals are more favourable toward foreign migrants than undereducated ones.

Abstract *In questo articolo studiamo la relazione tra disallineamento fra formazione e lavoro e atteggiamento verso la migrazione. Il disallineamento tra formazione e lavoro si osserva quando un posto di lavoro è occupato da un individuo con un livello di istruzione diverso da quello richiesto dal posto stesso. Gli individui sono definiti "overeducated" se in possesso di un titolo di studio superiore a quello richiesto dalla loro occupazione. Al contrario sono classificati come "undereducated" se in possesso di un livello di studio inferiore a quello richiesto. Usando dati del round 7 della European Social Survey (ESS), questo paper suffraga l'ipotesi che gli individui con qualifica superiore a quella richiesta dalla professione siano più favorevoli ai migranti.*

Key words: Attitudes towards migrants, Educational mismatch, Overeducation, Undereducation, Multilevel models.

¹ Economics and Political Science Area, INSEAD, Paris

² Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan

1 Introduction

Concerns about immigration can originate from economic origins as well as cultural and social reasons (Card, 2005). The main theoretical background behind economic reasons is the Group Conflict Theory defined by Blalock (1967) and Olzak (2005): negative outgroup sentiments are seen as defensive reactions to perceived intergroup competition for scarce goods. A competition that manifests itself in the labor market as a fear towards migrants with the same set of skills, and in the welfare system, as low skill (poor) immigrants can be perceived as a burden for public finance by richer natives. At the individual level, there is considerable empirical evidence about the fact that anti-immigration attitudes are more frequent among those with a lower level of education, skills and income (Coenders and Scheepers, 2003). Among the sociological and cultural theories, the most popular one is ethnocentrism, i.e. technical name for the view of things in which one's own group is the center of everything, and all others are scaled and rated with reference to it. Consequently, ethnocentric people perceive their social group as the center and judge other groups based on the different preferences with their own group. Manevska and Achterberg (2011) proved that ethnocentrism increase the sense of hostility towards migrants. In light of the results found in literature, these social and cultural factors seem to play a crucial role in explaining anti-immigration attitudes. Anyhow, the empirical analysis of these determinants is complicated, because the relationship between attitudes and these psychological characteristics can derive from reverse causality, rising problems of endogeneity. In this study, we focus on labour market competition and attitudes towards migrants. From a theoretical standpoint, standard economic models focus on the effect of immigration on the different factors of production. (Hainmueller and Hiscox, 2007). Assuming full employment, wage flexibility, a relative low skill level of immigrants compared to natives and perfect substitutability between migrants and natives in each skill category, the effect of immigrants entering the labor supply is a decrease in the wages of native low-skilled workers. This happens because low-skilled labour is applied to fixed amounts of the other factors, which implies that the real wages of the lower skilled will decline while the earnings of owners of land, capital, and skills will increase. This model is often referred to as “factor-proportions” analysis (Borjas, 1999) and it predicts that local low skilled workers are the real losers from immigration, and hence reasonably the category more threatful. Several studies have found strong evidence that concerns about labour market competition are a relevant determinant of attitudes toward immigration both in Europe and US. Most of these studies draw upon the evidence that there is a strong positive correlation between the skill levels of respondents (as measured by education levels) and their support for immigration. Hence, this may suggest that low skilled (less educated) workers are afraid of being forced to compete for jobs with immigrants, which are normally low-skilled (Scheve and Slaughter, 2001).

In this study, we propose the educational mismatch as a proxy for labor market competition. Educational mismatch occurs when the required level of education for a particular job diverges from the employee's attained level of education. The level

Educational mismatch and attitudes towards migration in Europa

of attained education could be higher than needed for the job, in which case the worker is over-educated, or lower than required, in which case the worker is under-educated.

Our prediction is that - across cells of occupation and skills - overeducated workers are more supportive towards immigration with respect to matched and undereducated workers. The theoretical hypothesis behind this is that they may feel more secure and less substitutable by migrant workers due to their overqualification for that specific job. On the contrary, undereducated and matched workers should feel more in danger, particularly the former.

After this introduction the paper is organized as follows: section 2 presents the database used and the statistical model adopted in the analysis. Section 3 displays the results of the regression analysis while section 4 concludes.

2 Data and Methods

We employ data from round 7 of the ESS (2015). The ESS is an academically driven cross-national survey which issues a multidimensional questionnaire across several European countries every two years. In the seventh round (the most recent released), there were 21 countries surveyed: Austria, Belgium, Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Ireland, Israel, Lithuania, Netherlands, Norway, Poland, Portugal, Slovenia, Spain, Sweden, Switzerland and the United Kingdom. Round 7 includes rotating modules on “Attitudes towards immigration and their antecedents”, which we exploit. As said in the introduction, the focus of our study is on the labour market competition, thus we need a measure of the realistic threat. ESS round 7 provides three questions to measure it:

- Would you say that people who come to live here generally take jobs away from workers in [country], or generally help to create new jobs?
- Most people who come to live here work and pay taxes. They also use health and welfare services. On balance, do you think people who come here take out more than they put in or put in more than they take out?
- Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries?

All three questions are measured in a scale from 0 to 10. In order to measure attitudes toward migrants we build an index computing the mean between the items mentioned above (an exploratory factor analysis stated that only one latent factor explain the scores of the three items). As far as the educational mismatch is concerned, there are several methods to measure the required level of education, and thus the extent of over/under education: job analysis (Green, Kler, and Leevs, 2007), worker self-assessment (Alba-Ramirez, 1993) and realized matches (Bauer, 2002). The first is the most “objective” measure, as it is based on the information contained in occupational classification documents which can be translated directly into number of years of schooling from 0 to 18. The second is instead the most

“subjective” measure, as workers are asked about the required education level of their job. The latter method entails two very similar approach: one, proposed by Verdugo and Verdugo (1989), uses the mean level of schooling of those working in the same occupations. It follows that any workers whose educational level is at least one standard deviation above the mean are deemed overeducated; those with one standard deviation below the mean are considered undereducated. In this paper we use the approach proposed by Kiker, Santos, and De Oliveira (1997), which is a variant of Verdugo and Verdugo (1989) in that it uses the mode the acquired schooling for workers in the same occupation and does not use the two standard deviations interval around the centralized measure. Workers with education level more or less than the modal value are considered over/under educated, respectively. In this framework, for every country in the dataset, we divide workers in the 9 major ISCO groups¹ and in three main categories of education. It follows that workers with education level more or less than the modal value of their occupation in their country are considered over (OE) / under (UE) educated, respectively. Finally, individuals with an educational level equal to the mode are defined as matched (M). After having defined overeducated individual, undereducated and matched ones we run a multilevel linear model to address whether attitudes towards migrants depend on over/undereducation. In order to establish in a more satisfying way the determinants of attitudes and to take into account the hierarchical structure of the ESS data by considering not only the individual characteristics (first level variables), but also the contextual conditions we use a two-level regression model. The model is specified as follows:

$$y_{ij} = \beta_0 + \sum_{h=1}^k \beta_h x_{hij} + \xi_{ij}$$

where y_{ij} is the dependent variable representing the attitude of individual i placed in the country j . x_{hij} are also all the other covariates such as gender (males, females), the country of birth (foreign or less), the age of respondents, the size of municipality and the level of education (less than upper secondary, secondary and tertiary). Overeducation (OE), Undereducation (UE) and matched education (M) are obviously included in the model. ξ_{ij} stands for the error term, which is decomposed as follows.

$$\xi_{ij} \equiv \zeta_j + \epsilon_{ij}$$

As depicted in the last equation, ξ_{ij} are defined as the sum of the first and second level errors. The random intercept ζ_j is a country-specific parameter that captures heterogeneity at country level. ϵ_{ij} is the residual component measured at country and individual level.

¹ The ISCO groups are 1) managers, 2) professionals, 3) technicians and associate professionals, 4) clerical support workers, 5) service and sales workers, 6) skilled agricultural, forestry and fishery workers, 7) craft and related trades workers, 8) plant and machine operators, and assemblers, 9) elementary occupations. We dropped ISCO group 6 due to a very low number of observations at the country level

3 Results

Results of the empirical analysis are presented in Table 1 in two different specification, including the effects of OE, UE and M conditions.

Table 1: Results for an OLS regression to study the determinants of attitudes towards migrants

<i>Explanatory variables</i>	<i>Model 1</i>	<i>Model 2</i>
Age	0.004	0.004
Age ²	-0.001	-0.001
<i>Gender</i>		
Males (Ref.)		
Females	-0.118***	-0.118***
<i>Born in the country</i>		
Yes (Ref.)		
No	0.449***	0.449***
<i>Size of Municipality</i>		
Big City (Ref.)		
Suburbs of big city	-0.069**	-0.069**
Town or Small city	-0.078***	-0.078***
Country Village	-0.145***	-0.145***
Home in Countryside	-0.147***	-0.147***
<i>Education</i>		
< Upper Secondary (Ref.)		
Upper Secondary	0.042	0.042
Tertiary	0.175**	0.175**
<i>Education – Job matching</i>		
Overeducated (OE)	0.187***	
Matched (M)	0.072*	-0.115***
Undereducated (UE)		-0.187***

* significant at 10% level; ** significant at 5% level; *** significant at 1% level.

As expected individuals with higher level of education are more favourable towards immigrants than lower educated as showed in Models 1 and 2. The introduction of Overeducation/Undereducation dummies support the hypotheses of the contribution: overeducated individuals have more positive attitudes towards migrants than matched and undereducated ones. This finding is line with the assumption of the presence of higher competition in the labour market for individuals with lower education. In addition, this competition has been made stronger if the individual is employed in a workplace that requires higher than those possessed by the worker.

With respect of control variables results are in line with expectations: people living in big cities are less opposite to migrants than those living in suburbs or small city. In addition, individuals that were born in foreign countries are more likely to display positive attitudes towards migrations.

4 Conclusions

This paper implemented a multilevel linear model in order to test the hypothesis that overeducated individuals are more likely to show more favourable attitudes towards foreign immigrants. In order to test the assumption, data from the round 7 of the ESS were used. Results support the hypothesis that overeducated individuals are more likely to be positively oriented towards foreign immigrants. This is in line with the present literature which underlines as the effect of immigrants entering the labour supply increases the perception of the risk of becoming unemployed for the less skilled natives.

References

1. Alba-Ramirez, A.: Mismatch in the Spanish labor market: overeducation? *Journal of Human Resources*, 28(2), 259-278 (1993).
2. Bauer, T.K.: Educational mismatch and wages: a panel analysis. In: *Economics of Education Review*, 21(3), 221-229 (2002).
3. Blalock, H. M.: *Toward a theory of minority-group relations*. New York: Wiley, Chicago (1967)
4. Borjas, G. J.: The economic analysis of immigration. In: *Handbook of labor economics*. Elsevier, 1697-1760 (1999).
5. Card, D.: Is the new immigration really so bad? *The Economic Journal*, 115.507, F300 (2005)
6. Coenders, M., Scheepers P.: The effect of education on nationalism and ethnic exclusionism: An international comparison. *Political psychology* 24(2), 313-343 (2003).
7. ESS: Round 7 Module on Attitudes towards Immigration and their Antecedents - Question Design Final Module in Template. In: London: Centre for Comparative Social Surveys (2015).
8. Green, C., Kler, P., Leeves, G.: Immigrant overeducation: Evidence from recent arrivals to Australia". In: *Economics of Education Review* 26(4), 420-432 (2007).
9. Hainmueller, J., Hiscox, M.J.: Educated preferences: Explaining attitudes toward immigration in Europe. *International organization* 61(2),399-442 (2007).
10. Kiker, B.F., Santos M.C, De Oliveira M.M.: Overeducation and undereducation: evidence for Portugal. *Economics of Education Review* 16(2), 111-125 (1997).
11. Manevska, K., Achterberg P.: Immigration and perceived ethnic threat: Cultural capital and economic explanations. *European Sociological Review* 29(3), 437-449 (2011).
12. Olzak, S.: *The dynamics of ethnic competition and conflict*. Stanford University Press, Stanford (2005).
13. Scheve, K.F., Slaughter, M.J.: Labor market competition and individual preferences over immigration policy. *Review of Economics and Statistics* 83(1), 133-145 (2001).
14. Verdugo, R.R., Verdugo, N. T.: The impact of surplus schooling on earnings: Some additional findings. *Journal of Human Resources*, 24(4), 629-643 (1989).

Soft thresholding Bayesian variable selection for compositional data analysis

Selezione di Variabili Bayesiana con funzioni di soglia per l'analisi di dati di composizione

Matteo Pedone, Francesco C. Stingo

Abstract

We investigate a regression model for the analysis of compositional data based on thresholding functions. The proposed soft thresholding mechanism induces sparsity in the regression model and selects only relevant covariates. We develop a Metropolis-Adjusted Langevin Algorithm (MALA) algorithm to efficiently explore the model space. First, we present our modeling approach within the linear regression framework and then extend it to the Dirichlet-Multinomial regression framework.

Abstract *Proponiamo un metodo generale per l'analisi di dati di composizione del microbioma al fine di stabilire la loro relazione con covariate esterne. Per indurre sparsità nel modello di regressione e per selezionare solo le associazioni più rilevanti abbiamo adottato un metodo di selezione delle variabili basato su una funzione di soglia. Inoltre, abbiamo implementato l'algoritmo Metropolis-Adjusted Langevin Algorithm (MALA) per rendere più efficace l'esplorazione dello spazio parametrico. Tale metodo è prima presentato nel contesto della regressione lineare per poi essere esteso alla regressione Dirichlet-Multinomial.*

Key words: Bayesian Variables Selection, Dirichlet-Multinomial, Microbiome data, Metropolis-Adjusted Langevin Algorithm, thresholding function

Matteo Pedone

Università degli Studi di Firenze, Dipartimento di Statistica, Informatica, Applicazioni
viale Morgagni, 59 - 50134 Firenze,
e-mail: matteo.pedone@unifi.it

Francesco C. Stingo

Università degli Studi di Firenze, Dipartimento di Statistica, Informatica, Applicazioni
viale Morgagni, 59 - 50134 Firenze,
e-mail: francescoclaudio.stingo@unifi.it

1 Introduction

In the context of bio-medical studies, recent developments of sequencing technology have resulted rich and complex datasets that may contain information crucial to the etiology, diagnosis, prognosis, and treatment of a large variety of diseases [6]. The analysis of these high-dimensional datasets requires the development of new methods; specifically, we are interested in the analysis of compositional data, such as the case of studies that focus on the microbiome. In particular, the relationship between human microbiome and clinical covariates or enviromental predictors can provide insights into personalized strategies in the prevention of several diseases. As the microbial data are high-dimensional and present several peculiarities (uneven sampling depth, overdispersion and zero inflation) they pose a severe challenge from both a methodological and a computational perspective. In such a context variable selection methods play a crucial role as they can identify significant association between sets of covariates and taxa of the microbiome.

Our goal is twofold: first, we explore a soft thresholding approach for variable selection in the context of compositional data, such as the microbial counts, we then focus on the computational aspects, and design a MCMC algorithm with Metropolis-Adjusted Metropolis Algorithm (MALA) to speed up the exploration of posterior density of the regression coefficients. The manuscript is organized as follows: first, introduce MALA, and describe how it can be used for a thresholding-based method for linear regression; its variable selection performances are investigated in a small simulation study, see Section 2. In Section 3 we propose a Bayesian Dirichlet-Multinomial regression model for the analysis of microbiome abundance tables.

2 Introduction to Metropolis-Adjusted Langevin Algorithm

Let \mathcal{X} be an open subset of \mathbb{R}^d with positive and continuously differentiable density π , with respect to a Lebesgue measure. The Langevin diffusion $\{x_t, t \geq 0\}$ associated with π is the solution of the following stochastic differential equation:

$$dx_t = \frac{1}{2} \Sigma \nabla \log \pi(x_t) dt + \Sigma^{1/2} dW_t, \quad (1)$$

where $\{W_t, t \geq 0\}$ is a standard d -dimensional Brownian motion, and Σ is a given positive definite symmetric matrix. Under appropriate assumptions on π [3], it can be shown that the dynamic generated by (1) is ergodic with unique invariant distribution π , it is hence possible to sample from the invariant distribution π . As only in few cases (1) allows an analytical solution, a discretization approach is adopted, even though a bias is introduced (π is not invariant with respect to the Markov chain defined by the discretization). Using the Metropolis-Hastings algorithm, a Markov chain $\{x_t, t \in \mathbb{N}\}$ is constructed, where at each step x^* is generated from a proposal density $q(x_t, \cdot)$. The candidate is then accepted with probability

$\alpha(x, x^*) = 1 \wedge \pi(x^*)q(x^*, x)/\pi(x)q(x, x^*)$. The Euler-Maruyama discretization for (1) leads to the Metropolis-Adjusted Langevin Algorithm (MALA), in which the proposal distribution is defined as:

$$x^* = x + (h/2)\Sigma\nabla\log\pi(x) + \sqrt{h}\Sigma^{1/2}\varepsilon, \quad (2)$$

where ε is a standard Gaussian random variable in \mathbb{R}^d . Note that MALA requires the evaluation of the target function's gradient in the definition of the proposal mechanism. This proposal density results in chains that quickly move toward *far but more likely regions* of the parameter space [2].

2.1 MALA for linear regression model

First, we test a MALA algorithm for multiple linear regression models. We are interested in high-dimensional data, with the number of covariates being greater than the number of observations ($n \ll p$). Our regression model is based on a thresholding function that induces sparsity and allows us to select covariates, setting to zero those coefficients whose effect is smaller than a stochastic threshold, hence eliminating those covariates whose association with the response variable is not strong. The approach we follow is inspired by the work of [5], where the threshold is interpreted as minimum effect size. Randomness in the thresholding function is characterized by its prior distribution; this modeling strategy implies that both the magnitude but the variability of the regression coefficients are taken into account in the selection mechanism. We assume the following data generating model: $\mathbf{Y} \mid \mathbf{X}, \beta, \sigma^2 \sim N(\mathbf{X}^T \beta, \sigma^2 \mathbf{I}_p)$, where \mathbf{Y} is a n -dimensional column vector, \mathbf{X} is a $n \times p$ matrix of covariates and \mathbf{I} is the $p \times p$ identity matrix. The Bayesian thresholding function is such that the vector β is function of the p -dimensional vector θ and the scalar t : $\beta = \theta \mathbb{1}_{\|\theta\| > t}$, for $t \geq 0$, $t \sim U(a_t, b_t)$. The priors for θ and σ^2 are respectively $\theta \sim N(\theta_0, \Sigma_0)$, $\sigma^2 \sim IG(a_\sigma, b_\sigma)$.

In order to estimate the model an MCMC algorithm is used where the main interest is the posterior distribution of the regression coefficient $p(\theta \mid \cdot)$:

- Update θ (MALA): given current θ draw $\theta^* \sim N(\mu(\theta), \delta_\beta \mathbf{I})$; set $\beta^* = \theta^* \mathbb{1}_{\|\theta^*\| > t}$.
- Update σ^2 (MH): given current σ^2 , draw $\sigma^{2*} \sim IG(\delta_\sigma, (\delta_\sigma - 1)/\sigma^2)$.
- Update t (MH): given current t , draw $t^* \sim U(t - \delta_t, t + \delta_t)$.

MALA, that is used in the θ s upgrade step, adds a drift on the proposal distribution -obtained by the study of the gradient of the posterior distribution at the current state of the Markov chain- in order to hit the higher probability region of the space of the parameters. In order to implement MALA, we exploit the following smooth approximation [1]:

$$\mathbb{1}_{\|\theta\| > t} \simeq \frac{1}{2} \left\{ 1 + \frac{2}{\pi} \arctan\left(\frac{\theta^2 - t^2}{\varepsilon_0}\right) \right\} \text{ for } \varepsilon_0 \rightarrow 0 \quad (3)$$

In terms of hyperparameter setting for the parameters θ, t , we can use the following equality to set the appropriate *a priori* level of sparsity [5]:

$$\mathbb{E}_t[Pr(|\theta| > t | t)] = \int_{\mathcal{T}} Pr(|\theta| > t | t) p(t) dt. \quad (4)$$

2.1.1 Simulation Study

We run a small simulation study aimed at investigating the performances of the outlined method. The companion R/C code is available online at <https://github.com/mattpedone/tmala>. We ran 50,000 MCMC iterations, retaining one every 50 observations. The results are based on 20 repetitions; standard deviation is given within parenthesis. The $n \times p$ covariates matrix \mathbf{X} is simulated by sampling from a $p - 1$ dimensional multivariate normal, that is $X \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ and then a n -dimensional vector of 1s is placed as first column. \mathbf{Y} , the n -dimensional response variable vector is obtained as a linear combination of $p^* \ll p$ columns of \mathbf{X} , plus an error term is added. We set $n = 50$, $p = 100$, $p^* = 2$, the regression coefficient is $\beta = \{\beta_1, \dots, \beta_p\} = \{1, 3, 0, 0, 3, 0, \dots, 0\}$, $\mathbf{y}_i = \mathbf{x}_i^T \beta + \varepsilon_i$, $\varepsilon_i \sim N(0, 1)$. It took 339.663 seconds to run 20 50,000-iteration Chains on a Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz CPU machine. The results are reported in Table 1.

TPR	MCC	AUC	MSE
0.9167	0.8921	0.9558	3.0153
(0.0331)	(0.0409)	(0.0175)	(0.852)

Table 1: True Positive Rate, Matthew Correlation Coefficient, Area Under the Curve and Mean Square Error for the prediction of $\hat{\mathbf{y}}$.

3 A threshold approach for Dirichlet Multinomial regression

In this section we introduce a regression model that can be used for the analysis of compositional data. Let $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ})$, $i = 1, \dots, n$, $j = 1, \dots, J$ ($\mathbf{Y} \in \mathbb{R}^{n \times J}$) be the vector of counts representing, for example, the taxonomic abundance table obtained from the i -th patient, with y_{ij} being the i -th frequency of the j -th microbial taxon. Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ ($\mathbf{X} \in \mathbb{R}^{n \times P}$) be the $n \times P$ matrix of external covariates, where $p = 1, \dots, P$ and P is the number of external covariates. The taxonomic counts may be modeled with a Multinomial distribution [7]: $\mathbf{y}_i | \phi_i \sim \text{Multinomial}(|\mathbf{y}_i|, \phi_i)$, with $|\mathbf{y}_i| = \sum_{j=1}^J y_{ij}$ being the sum of all the counts in the vector, and where the J -dimensional ϕ parameter vector is defined on the J -dimensional simplex: $\mathcal{S}^{J-1} = \{(\phi_1, \dots, \phi_J) : \phi_j \geq 0, \forall j, \sum_{j=1}^J \phi_j = 1\}$. We impose a conjugate Dirichlet prior on ϕ_i , $\phi_i \sim D(\gamma)$, $\gamma = (\gamma_1, \dots, \gamma_J)$, $\gamma_j > 0 \forall j$, where γ is the J -dimensional

vector of parameters. Exploiting conjugacy, ϕ can be integrated out, obtaining the compound Dirichlet-Multinomial (DM) model [4]: $\mathbf{y}_i \sim DM(\gamma)$ with probability mass function

$$f(\mathbf{y} | \gamma) = \frac{\Gamma(|\mathbf{y}| + 1)\Gamma(|\gamma|)}{\Gamma(|\mathbf{y}| + |\gamma|)} \times \prod_{j=1}^J \frac{\mathbf{y}_j + |\gamma|}{\Gamma(\gamma_j)\Gamma(\mathbf{y}_j + 1)} = \prod_{i=1}^n \binom{|\mathbf{y}_i|}{\mathbf{y}_i} \frac{\prod_{j=1}^J (\gamma_j)^{(\mathbf{y}_j)}}{(|\gamma|)^{(|\mathbf{y}_i|)}}.$$

where $|\gamma| = \sum_j \gamma_j$ and $(a)_{(k)} = a(a+1)\dots(a+k-1)$ denotes the rising factorial. Note that this is a much more flexible model than a simple Multinomial model as it can take into account overdispersion.

Next, external covariates are embedded into the model through a log-linear regression model, imposing dependence of the DM parameters on external covariates in \mathbf{X} . In particular, let $\zeta_j = \log(\gamma_j)$ and assume that $\zeta_j = \alpha_j + \sum_{p=1}^P \beta_{pj} \mathbf{x}_p$. The intercept α_j , $\alpha \in \mathbb{R}^J$, is the log baseline parameter for the j -th taxon, for whom we assume normal prior distribution, and the regression parameter vector β_j , $\beta \in \mathbb{B}^{P \times J}$, $\mathbf{B} = (\beta_1, \dots, \beta_J) \in \mathbb{R}^{P \times J}$, is the effect of the external covariates on the j -microbial taxa counts. Particularly when the number of external covariates is greater than the number of patients available, variable selection procedures are likely to be exploited in order to take into account only relevant association between taxa and covariates. In this regard the soft thresholding variables selection method is adopted in order to detect non-zero β_{pj} . Note that β is just a function of θ and t , that are two random variables, $\theta \in \Theta^{P \times J}$, $t \in \mathbb{R}$.

Hence the model on the external covariates can be summarized similarly as before: $\alpha_j \sim N(0, \sigma_j^2)$, $\beta_{pj} = \theta_{pj} \mathbb{1}_{[|\theta_{pj}| > t]}$, for $t \geq 0$, $\theta_j \sim N_p(\mu_0, \Sigma_0)$ and $t \sim U(a_t, b_t)$. The log-likelihood $\ell_n(\mathbf{y} | \cdot)$ takes the form:

$$\begin{aligned} \ell_n(\mathbf{y} | \cdot) &= \sum_{i=1}^n \sum_{j=1}^J \sum_{k=0}^{y_{ij}-1} \log(\exp\{\mathbf{x}_i^T \beta_j\} + k) - \sum_{i=1}^n \sum_{k=0}^{|\mathbf{y}_i|-1} \log\left(\sum_{j=1}^J \exp\{\mathbf{x}_i^T \beta_j\} + k\right) \\ &\quad + \sum_{i=1}^n \log\left(\binom{|\mathbf{y}_i|}{\mathbf{y}_i}\right). \end{aligned} \tag{5}$$

3.1 MCMC algorithm for posterior computation

In order to approximate our posterior density of interest, $p(\theta | \cdot)$, we are currently working on the development of a Markov Chain Monte Carlo (MCMC) algorithm, based on a MALA step, aimed at an efficient exploration of the model space.

Being s the index for MCMC iteration, $s \in \mathbb{N}$, denote by $L(\mathbf{y} | \alpha, \theta, \mathbf{X}, t)$ the likelihood function, the key steps in our posterior computation algorithm include, for a general $(s+1)$ -th step:

- Update α_j (MH): given current α_j draw $\alpha_j^* \sim N(\alpha_j, \delta_\alpha)$ for $j = 1, \dots, J$.

- Update θ_j (MALA): given current θ_j draw $\theta_j^* \sim N_p(\mu(\theta_j), \delta_\beta)$, for $j = 1, \dots, J$; set $\beta_{pj}^* = \theta_{pj}^* \mathbb{1}_{[|\theta_{pj}^*| > t]}$.
- Update t (MH): given current t , draw $t^* \sim U(t - \delta_t, t + \delta_t)$.

where δ_α , δ_β , δ_t are the tuning parameters for the proposals' distribution and $\mu(\theta_j) = \theta_j^s + \frac{\delta_\beta^2}{2} \left(\nabla_{\theta_j^s} \log p(\beta_j^s | \cdot) \right)$ (note that it is an adaptation of (2)).

4 Discussion

We presented a general framework for variable selection with a Bayesian thresholding mechanism that effectively exploits MALA. Our aim is to develop a computationally efficient approach for the selection of relevant associations between microbiome taxa and clinical and environmental covariates. We are aware of some limitations of the proposed approach; for example, one of the main MALA's disadvantages is its sensitivity to tuning parameters. In particular, tuning parameters should be set to achieve an optimal convergence rate, but it is well-established that MALA's optimal convergence rate is different in the transient compared to in the stationary phase of the Markov Chain. Hence, we are currently working on the implementation of adaptive versions of MALA that can effective beyond the linear regression model.

References

1. Cai, Q., Kang, J. Yu, T.: Bayesian network marker selection via the thresholded graph Laplacian Gaussian prior. *Bayesian Analysis* (2018)
2. Girolami, M. and Calderhead, B.: Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 123–214 (2011)
3. Khasminskii, Rafail: *Stochastic stability of differential equations*. Springer Science & Business Media, (2011)
4. Mosimann, J. E.: On the compound multinomial distribution, the multivariate β -distribution, and correlations among proportions. *Biometrika* **49**, 65–82 (1962)
5. Ni, Yang and Stingo, Francesco C and Baladandayuthapani, Veerabhadran: Bayesian graphical regression. *Journal of the American Statistical Association* (2018)
6. Tang, Yunfan and Ma, Li and Nicolae, Dan L.: A phylogenetic scan test on a dirichlet-tree multinomial model for microbiome data. *The Annals of Applied Statistics* **18**,1 (2018)
7. Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., Vanucci, M.: An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics* **18**, 94 (2017)

Sentiment-driven investment strategies: a practical example of AI-powered engines in a corporate setting

Strategie d'investimento guidate dal sentiment: un esempio pratico di Intelligenza Artificiale in contesto aziendale

Mattia Pedrini, Sebastian Donoso, Enrico Deusebio, Nicola Donelli, Gabriele Arici, Andrea Cosentini
Paola Mosconi, Diego Ostinelli and Claudio Cocchis

Abstract In this paper we present an investment strategy on the S&P500 driven by a measure of sentiment from both Twitter and Reuters news. The sentiment from Twitter is extracted using a custom Deep Neural Network while the sentiment from Reuters news is provided from "Reuters Machine Readable News" service. We use sentiment indices as regressors for a number of state-space models with intraday returns observations and combine these models in order to produce a single forecast of the index return. We finally build our strategy deriving a buy/stay/sell signal from our forecast and adding a stop loss threshold.

Abstract In questo articolo presentiamo una strategia di investimento sull'S&P500 guidata esclusivamente da una misura del sentiment ottenuta da Twitter e delle notizie Reuters. Il sentiment di Twitter viene estratto utilizzando una Deep Neural Network creata ad hoc mentre il sentiment delle notizie Reuters fornito dal servizio "Reuters Machine Readable News". Usiamo gli indici di sentiment come regressori per diversi modelli state-space aventi come osservazioni i rendimenti intraday dell'indice e combiniamo questi modelli in modo da produrre un'unica previsione dei rendimenti futuri dell'indice. Infine costruiamo la nostra strategia derivando un segnale buy/stay/sell dai rendimenti previsti ed aggiungendo una stop-loss.

Key words: sentiment analysis, deep learning, trading strategies, state-space models

Mattia Pedrini, Sebastian Donoso, Enrico Deusebio, Nicola Donelli, Gabriele Arici, Andrea Cosentini
CGnal S.r.l., Corso Venezia 43, Milano, e-mail: {mpedrini, sdonoso, edeusebio, ndonelli, garici, acosentini, datascience} @cgnal.com

Paola Mosconi, Diego Ostinelli, Claudio Cocchis
Banca IMI, Largo Mattioli 3, Milano, e-mail: {paola.mosconi, diego.ostinelli, claudio.cocchis} @bancaimi.com

1 Introduction

Literature on the relationship between investor sentiment and market behavior has been growing very fast in the last ten years. Particularly relevant for our work had been [2] that examined the relation between sentiment and returns of companies, [1] that used Deep Learning techniques to analyze investor sentiment from financial news, [3] that used text mining techniques and decision trees to develop a full stack trading support system based on micro-blogging messages and [4] that investigated the application of multivariate Hawkes process model to analyze the relation between sentiment of investors and market returns.

In this paper, we present a real-time alternative-data-driven engine developed to support a Trading Desk of IMI Investment Bank. The engine aggregates in real-time information from Twitter and Reuters News to produce sentiment indexes about the S&P500 and uses them to compute an advice signal to support daily and intraday trading order execution. The analytical engine is built combining a mixture of artificial intelligence algorithms, such as Deep Neural Network (DNN), for Natural Language Processing and Sentiment Estimation, and state-space econometric models, to produce returns forecasts. The engine, actually used by IMI Bank since September 2018, has significantly outperformed the basic buy-and-hold strategy.

The paper is organized as follows: in Section 2, we describe the analytical pipeline and the data used for training and testing it and in section 3 we assess the performance of our system against benchmark buy-and-hold strategy.

2 Methodology

We now review the structure of our Sentiment Driven Investment Strategy (SDIS), describing its building blocks and the data we used for our application.

2.1 Twitter Sentiment Extraction

We construct a Deep Neural Network (DNN) that associates each tweet with a Sentiment Score in $[0, 1]$ (1 is a strongly positive sentiment and 0 a strongly negative sentiment). The DNN we constructed is constituted by the blocks described below and implemented using Keras.

The Embedding Layer To map plain text from the Tweets into a numeric feature space, we start with an *Embedding Layer* based on a Word Embedding Matrix (WEM) built with the Global Vectors for Word Representation (GloVe) algorithm by the NLP group at Stanford University, described in [5]. This particular WEM embeds a dictionary of 10,000 words in \mathbb{R}^{100} .

ConvNet Layers The output representation from GloVe is followed by two Convolutional Layers [6] with the so-called *ConvNet* architecture. Each of our ConvNet Layers is a sequence of: convolutional layer, activation layer with Exponential Linear Unit (ELU) activation function [7], MaxPooling layer and Dropout layer to reduce overfitting.

Bidirectional LSTM Layers The next layers are devoted to exploit the sequential property of the spoken language through Long short-Term Memory (LSTM) units [8]. In particular we use Bidirectional LSTM layers, which had been shown by many authors (see [9], [10] and [11]) to be of great usefulness in the context of speech processing, because of their ability to process sequences of words together with their inverses. Our architecture is completed with two Bidirectional LSTM layers together with ELU activation layers (see Paragraph above) and concluded with a fully connected layer (Multilayer Perceptron).

Training Data and Model Validation We trained our DNN for the specific task of understanding sentiment from Tweets using the dataset described in [12], that contains 1,583,691 sentiment-labeled tweets.

The described architecture was selected after model validation, on the same dataset, minimizing Cross Entropy

$$\mathcal{L} = -\frac{1}{n} \sum_{i=1}^n y(x_i) \log K(x_i) + (1 - y(x_i)) \ln(1 - K(x_i)). \quad (1)$$

where $y(x_i)$ is the true label of tweet x_i and $K(x_i)$ is the output of the neural network given input tweet x_i .

2.2 Tendency Model

We call Tendency Model the macro-component of our algorithm that takes in input the sentiment scores extracted as in Section 2.1 and the sentiment scores from Thomson Reuters and outputs a *Tendency Signal*.

Daily Sentiment Indexes For both our sources separately, we define a mapping¹ between text inputs and stocks in the S&P500 index. For each stock, we take the hourly averages of the sentiment scores of the text inputs referring to it, obtaining two hourly scores for each stock. We then calculate two hourly scores for the whole index by taking a market-capitalization-weighted average of the stock scores from each source. Finally we average these hourly scores to get two Daily Sentiment Indexes, one for each source.

The state-space Model We model the intraday returns as a state-space model, taking the sentiment scores as exogenous regressors of for the space equation. Let n

¹ we use Cashtags and Hashtags for tweets and the AssetName field for Reuters News

be the sample size and let \mathbf{A}' denote the transpose of a matrix or vector \mathbf{A} . For $t \in \{1, \dots, n\}$ the model is:

$$\begin{cases} y_t = \mathbf{z}_t' \boldsymbol{\alpha}_t + \varepsilon_t \\ \boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \boldsymbol{\eta}_t \\ \boldsymbol{\alpha}_1 \sim N_2 \left(\begin{bmatrix} a \\ b \end{bmatrix}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_1} \right) \end{cases} \quad \begin{cases} \{\varepsilon_t\}_{t \geq 1} \stackrel{iid}{\sim} N(0, \sigma_r^2) \\ \{\boldsymbol{\eta}_t\}_{t \geq 1} \stackrel{iid}{\sim} N_2 \left(\mathbf{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \begin{bmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{22}^2 \end{bmatrix} \right) \end{cases}$$

where $y_t = \ln(p_t^c/p_t^o)$ is the logarithmic intraday return, p_t^o is the Open price, p_t^c is the Close price, $\mathbf{z}_t = [S_{t-1}, S_{t-2}]'$ is the vector of sentiment scores (from Twitter or Reuters) of the two previous days as defined in Section 2.1 and $\boldsymbol{\alpha}_t \in \mathbb{R}^2$ is the vector of states. The initial state mean $[a, b]'$ and its initial variance $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_1}$ are constants to be set. All the processes and initial variables are pairwise stochastically independent and the only parameters in the model are the variances of the innovations, $\boldsymbol{\theta} = (\sigma_r^2, \sigma_{11}^2, \sigma_{22}^2)$. To estimate these parameters we tried both Grid Search optimization using a given metric (Accuracy or Theoretical Market Return) or Maximum Likelihood Estimation, the best performances were obtained by combining the models obtained with those techniques into an Ensemble Model. Finally, we use the Kalman Filter for one-day-ahead forecast.

Investment Strategy To define an investment strategy from model forecasts we introduce a *positive* and a *negative confidence threshold* $\alpha_+, \alpha_- \in \mathbb{R}_{\geq 0}$ determining when the model output is too close to zero to be confident enough to operate on the market for a given day. Given a return forecast \hat{y}_t , the advised strategy b_t will be

$$b_t = f(\hat{y}_t) = \begin{cases} 1 \text{ or "BUY"} & \text{if } \hat{y}_t > \alpha_+ \\ 0 \text{ or "STAY"} & \text{if } -\alpha_- \leq \hat{y}_t \leq \alpha_+ \\ -1 \text{ or "SELL"} & \text{if } \hat{y}_t < -\alpha_- \end{cases} \quad (2)$$

All the operations are intended to be intraday². The confidence thresholds α_+, α_- are then determined optimizing the cumulative return of the strategy.

Ensembling The idea of using an ensembling strategy for our models came from back-testing the investment strategies: we realized that state space models with different hyperparameters tend to perform differently in different market conditions, in particular the combinations of hyperparameters will determine how quickly the model will react to changes in market direction, determining its momentum and how well will the model behave in high/low market volatility conditions.

The idea is then to identify the best model to be used in each market condition, and use that model to generate the tendency signal best suited to that market condition. The identification of the best model is done by ranking the models with respect to some chosen metric. The *Daily Adaptive Ensemble Model* will then work as follows.

² BUY means to buy when market opens and sell when market closes, SELL means to sell short when market opens and regulate the position when market closes

Given a performance metric M , let C_1, C_2, \dots, C_n be n selected investment strategies and relative state space models and k be the length of the time window over which we rank the models. Also, for every $i \in 1, \dots, n$ let $Y_{t,k}^i = \{y_s^i\}_{t-k \leq s < t} \subset \{-1, 0, 1\}$ be the set of Buy/Sell signals generated by the model C_i in the k days before a given day $t \in T$. Then the model used to generate the Buy/Sell signal for day t is $C_{\bar{i}}$, where

$$\bar{i} = \operatorname{argmax}_i \{M(Y_{t,k}^i)\}_{i=1, \dots, n} \quad (3)$$

if the metric M is a performance metric to be maximized. The model includes as base learners a family of ca. 8000 state space-models with varying parameters, and depends on a free parameter k and the chosen selection metric M , both to be optimized using historical data. The optimization led to $M = \text{Accuracy}$ and the resulting The Daily Adaptive Model have shown to consistently outperform the best base models.

3 Application

For our application, we use intraday returns of the S&P500 Index, Sentiment score from Thompson Reuters' News as a first measure of investors' sentiment and a stream of Tweets referring to assets in S&P500 as a second source of investors' sentiment. We consider the period from April, 4th, 2017 to March, 18th, 2019. We get the news sentiment data from Thomson Reuters News Analytics database³ and Twitter data from Twitter public APIs.

We back-tested our strategy, supported by a standard EWMA stop-loss, on the full extent of our time window but we think it is more interesting to present the actual performances of the SDIS, starting from when Banca IMI actually started using it for daily trading: thus the reference period goes from 7/9/2018 to 18/3/2019. The strategy is still running live and performances will be updated in the future.

Figure 1 shows the returns the SDIS and two benchmarks. We can clearly see that SDIS consistently outperforms the passive buy-and-hold strategy and the SPX intraday returns. Finally Table 1 shows the performances of SDIS in terms of a number of metrics of interest for traders traders.

Table 1 Performance metrics

Strategy	Annualized Returns	Volatility	Sharpe ratio	Sortino ratio	Downside Volatility
SDIS	+17,10%	11,71%	1,46	1,52	11,28%
BHIS	-2,63%	19,47%	-0,14	-0,18	14,63%
Intraday	-14,74%	15,00%	-0,98	-1,17	12,64%

³ For more details about Thomson Reuters News Analytics database please visit <https://financial.thomsonreuters.com/en/products/data-analytics/financial-news-feed/world-news-analysis.html>.

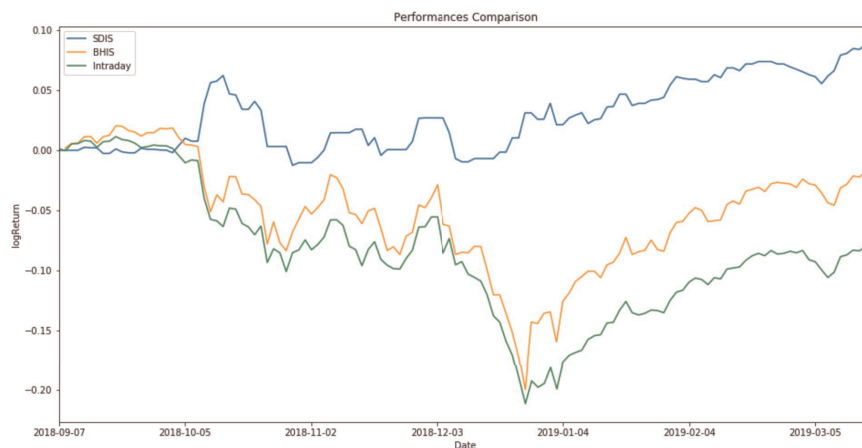


Fig. 1 Comparison of log returns from SDIS, Buy-and-hold strategy (BHIS) and intraday SPX

References

1. Day, M., Lee, C.: Deep learning for financial sentiment analysis on finance news providers. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1127–1134. (2016)
2. Baker, M., Wurgles, J.: Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance*. **61**, 1645–1680 (2006)
3. Nasser, A.A., Tucker, A., de Cesare, S.: Quantifying StockTwits semantic terms trading behavior in financial markets: An effective application of decision tree algorithms. *Expert Systems with Applications*. **42**, 9192–9210 (2015)
4. Yang, S., Liu, A., Chen, J., Hawkes, A.: Applications of a Multivariate Hawkes Process to Joint Modeling of Sentiment and Market Return Events. *SSRN Electronic Journal*. (2017) doi: 10.2139/ssrn.2954079
5. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. (2014)
6. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. **86**, 2278–2324 (1998)
7. Clevert, D., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289* (2015)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation*. **9**, 1735–1780 (1997)
9. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*. **18**, 602–610 (2005)
10. Liwicki, M., Graves, A., Fernandez, S., Bunke, H., Schmidhuber, J.: A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In: *Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007*. (2007)
11. Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*. **31**, 855–868 (2009)
12. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*. **1**, (2009)

Betting on football: a model to predict match outcomes

Scommettere sul calcio: un nuovo modello per prevedere l'esito delle partite

Marco Petretta, Lorenzo Schiavon and Jacopo Diquigiovanni

Abstract In the context of sports betting, estimates of probabilities related to specific events in a football match are quantities of deep interest. Starting from appropriate marginal and conditional distributions, a joint distribution is introduced in order to model the dependence between the number of goals scored by the two opposing teams. This model allows to obtain estimates of probabilities for all the possible match outcomes and, in so doing, to develop a betting strategy. An application is presented in this paper: it shows encouraging results and highlights satisfactory return in the long run.

Abstract Nell'ambito delle scommesse sportive, risulta di particolare interesse la stima delle probabilità associate a specifici eventi di una partita di calcio. A partire dalla definizione di opportune distribuzioni marginali e condizionate, viene introdotta una distribuzione congiunta capace di modellare la dipendenza tra il numero di gol segnati dalla squadra in casa e quelli segnati dalla squadra in trasferta. Questo modello consente di ottenere una stima delle probabilità dei diversi esiti di un incontro, da cui è possibile sviluppare una strategia di scommesse. Inoltre viene presentata un'applicazione che mostra risultati incoraggianti, con rendimenti soddisfacenti nel lungo periodo.

Key words: Betting strategy, Conditional distribution, Dixon and Coles model, Football (soccer), Marginal distribution, Poisson distribution

Marco Petretta
e-mail: marco.petretta1@gmail.com

Lorenzo Schiavon
Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, IT-35121 Padova, Italy, e-mail: lorenzo.schiavon@phd.unipd.it

Jacopo Diquigiovanni
Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, IT-35121 Padova, Italy, e-mail: jacopo.diquigiovanni@phd.unipd.it

1 Introduction

Moving from the paper written by Maher [6], several authors have proposed different approaches to model the final outcome of an association football (referred to simply as “football” hereafter) match in order to develop a profitable betting strategy (see, for example, [4]). A critical issue as regards this topic concerns the inclusion of a dependence structure between the number of goals scored by the two opposing teams, the importance of which has been emphasized, for example, by Dixon and Coles [1]. In view of this, the aim of this work is to develop a model (Mar-Co) that allows the number of goals scored by the two opposing teams to be dependent by the use of a specific joint probability mass function (PMF) based on marginal PMFs (Mar-) and conditional PMFs (-Co). The paper is organised as follows: in Sect. 2 the model is presented, in Sect. 3 it is used to develop a betting strategy and in Sect. 4 an overview of the main results and drawbacks is provided.

2 The model

The model is based on the Dixon and Coles approach [1], that is briefly summarised as follows. It considers the number of goals scored by the home team (away team) as a Poisson random variable whose expected value depends on the attack rate of the home team (away team) and on the defence rate of the away team (home team). The final ‘pseudolikelihood’ includes also a parameter which allows for the home effect, a dependence structure between the two variables when the final scores are 0-0, 1-0, 0-1, 1-1 and a weighting function ϕ that enables the changes in team performance during the season to be included. The crucial difference between our model and the Dixon and Coles one lies in the specification of a joint PMF that allows X , *i.e.* the number of goals scored by the home team, and Y , *i.e.* the number of goals scored by the away team, to be dependent also when the final results are different from those abovementioned. Let us focus, for the moment, on the case of independence between X and Y . According to the Dixon and Coles model, the marginal PMFs of X and Y can be defined as follows:

$$\begin{aligned} X_i &\sim \text{Poisson}(\lambda_{ij}), & \log(\lambda_{ij}) &= \nu + \alpha_i + \beta_j + \gamma \\ Y_j &\sim \text{Poisson}(\mu_{ji}), & \log(\mu_{ji}) &= \nu + \alpha_j + \beta_i, \end{aligned} \quad (1)$$

with ν intercept, α_i, β_i attack and defence parameters of team i such that $\sum_i \alpha_i = \sum_i \beta_i = 0$ and γ a parameter which allows for the home effect. Under the assumption of independence, the joint PMF is defined as the product of the two marginal PMFs.

In order to include a dependence structure in the model, an improvement consists of providing a possible conditional PMF $X|Y$ (or $Y|X$). A reasonable choice for $X|Y$ is the following:

$$X_i|Y_j = y \sim \text{Poisson}(\delta_X(\lambda_{ij}, y)), \quad \text{with} \\ \delta_X(\lambda, y) = \exp\{\theta_1 + \theta_2 \log \lambda + \theta_3 f(P(X > y))\},$$

with $\theta_1, \theta_2, \theta_3 \in \mathbb{R}$ and $f(\cdot)$ the *logit* function. In so doing, the mean of $X|Y$ varies according to the easiness with which a team, given his “marginal strength”, scores more goals than those scored in that specific match by the opposing team. Moreover, the variables are dependent when $\theta_3 \neq 0$ and the abovementioned independence case is verified when $\theta_1 = \theta_3 = 0$ and $\theta_2 = 1$. In particular, the latter case corresponds to the Dixon and Coles model when ρ , the parameter used by Dixon and Coles to allow the variables to be dependent, is fixed to 0.

Starting from the definition of joint PMF as product of conditional PMF and marginal PMF, we can define:

$$P_A(X_i = x, Y_j = y) = \frac{e^{-\delta_X(\lambda_{ij}, y)} \delta_X(\lambda_{ij}, y)^x}{x!} \frac{e^{-\mu_{ji}} \mu_{ji}^y}{y!} \quad (2)$$

It is immediate to verify that the marginal PMF of X obtained from the joint PMF (2) is no longer a Poisson PMF when $\theta_3 \neq 0$. In the same way, it is possible to define a different joint PMF by considering a conditional PMF of $Y|X$ and the marginal PMF of X proposed in Equation (1):

$$X_i \sim \text{Poisson}(\lambda_{ij}), \quad Y_j|X_i = x \sim \text{Poisson}(\delta_Y(\mu_{ji}, x)) \\ \delta_Y(\mu, x) = \exp\{\theta_1 + \theta_2 \log \mu + \theta_3 f(P(Y > x))\} \\ P_B(X_i = x, Y_j = y) = \frac{e^{-\delta_Y(\mu_{ji}, x)} \delta_Y(\mu_{ji}, x)^y}{y!} \frac{e^{-\lambda_{ij}} \lambda_{ij}^x}{x!} \quad (3)$$

Since we do not have reason to prefer one of the two formulations, we define the joint PMF of $(X_i = x, Y_j = y)$ as a mixture of P_A and P_B with weights equal to 0.5. Based purely on football considerations, the existence of two subpopulations appears to be a reasonable assumption: indeed, one is justified in expecting sometimes the home team to “react” to the performance of the away team, sometimes viceversa. From a statistical point of view, the marginal PMF of X (Y respectively) obtained from the mixture joint PMF is not a Poisson PMF, but it is more similar to a Poisson PMF than the marginal PMF obtained from P_A (P_B) since it is defined as the average of the non-Poisson marginal PMF obtained from P_A (P_B) and the Poisson marginal PMF obtained from P_B (P_A). Specifically, the lack of Poisson marginal PMFs is due to the desire to include a comprehensive dependence structure in the model in order to - hopefully - improve the overall predictive capability.

The estimates are obtained by a plug-in estimation process [3], a method commonly used when the number of parameters is high. Specifically, for each time point t , first of all $\hat{\nu}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}$ are obtained by using the Dixon and Coles estimation method with ρ fixed to 0. Secondly, $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ are obtained by maximizing the following function:

$$L_t(\theta, \hat{\nu}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}) = \prod_{k \in A_t} \frac{P_{A(k)}(\theta, \hat{\nu}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}) + P_{B(k)}(\theta, \hat{\nu}, \hat{\alpha}, \hat{\beta}, \hat{\gamma})}{2} \quad (4)$$

with $A_t = \{k : t_k < t\}$, t_k the time that match k is played and $P_{A(k)}$ ($P_{B(k)}$ respectively) the right side of Equation (2) (Equation (3)) with regard to match k . For the sake of simplicity, the weighting function ϕ is not included in Equation (4) since we believe θ not to have a relevant dynamic over time.

A one-step maximization of the pseudolikelihood (after including the weighting function ϕ proposed by Dixon and Coles) is also feasible and represents undoubtedly a more formal approach; nevertheless, empirical results suggest that the first estimation method, in addition to being computationally faster, provides estimations of the quantities of interest similar to those obtained by the second one.

3 Betting strategy and results

By implementing the model described in the previous section it is possible to obtain estimates of the probabilities p of each match outcome. In order to develop a betting strategy against odds provided by bookmakers, let us consider, for each time point t , the set $\mathcal{A}^{(t)} = \{a_i^t\}_i$, with $a_i^t = (e_i^t, m_i^t)$ the 2-tuple indicating money m_i^t bet on the event e_i^t . Specifically, the set $e^t = \{e_i^t\}_i$ is a subset of $E^{(t)}$, the set of all the possible outcomes ($I, X, 2$) of the matches played at time point t . The criteria to select the subset e^t is described as follows: first of all the probability p_i for each event i in $E^{(t)}$ is computed and each event, such that its probability is bigger than the inverse of the odds provided by at least one bookmaker j (i.e. $p_i > \frac{1}{o_{ij}}$), is taken into account. Secondly, each event whose probability p_i is much higher than the inverse of the average of the odds provided by different bookmakers ($p_i \gg \frac{1}{\bar{o}_i}$) is not included since this gap could be due to a lack of fundamental information (e.g. an injury of a key player during the last few days) in our model. Finally, each event whose probability is less than 0.1 is not included in order to develop a more conservative betting strategy.

The second fundamental step concerns the fraction of money allocated to each selected event (portfolio hereafter). The portfolio is chosen from a set of 10000 possible portfolios including both widespread portfolios (e.g. equally weighted portfolio, portfolio defined according to the Kelly criterion [5]) and randomly simulated portfolios. Thus, it is possible to compute, for each portfolio, the approximated distribution of gain G by simulation. In order to choose the *best* portfolio, the following five indicators are computed for each of the 10000 portfolios: expected value, variance, median, probability that G is greater than zero and 10% Value at Risk (VaR). These indicators are summarised by a unique index of performance that provides the favorite portfolio in terms of balance between return and risk. According to the abovementioned will to develop a fairly conservative strategy, the maximum amount

of money bet at each time point t is directly proportional to 10% VaR and always lower than 20% of the bankroll.

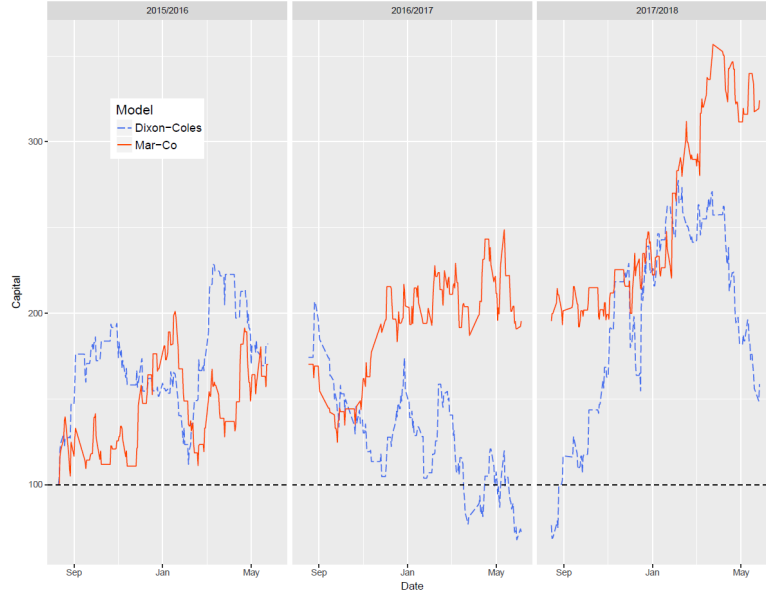


Fig. 1 Cumulative gain obtained by the Mar-Co model and the Dixon-Coles model from August 2015 to May 2018. The horizontal dashed line displays the starting capital.

In order to verify the performance of the method in betting markets, an application is provided as follows. The available data - downloaded from <http://www.football-data.co.uk/> - refers to the matches of the Italian Serie A, the Spanish La Liga, the German Bundesliga and the English Premier League 2015-2016, 2016-2017 and 2017-2018 seasons.

For each time point t , the Mar-Co model is estimated and, for each day on which at least one event is selected, the previously mentioned betting strategy is developed assuming a constant bankroll equal to 100 units. Moreover, the probabilities estimated by the Dixon and Coles model are obtained: in so doing it is possible to compare the results of the two approaches considering the same betting strategy. Figure 1 shows the cumulative gain of the two models: the Mar-Co model seems to achieve good results as the starting amount of money has more than tripled, reaching 324.14 units at the end of the three years. It is noticeable that the trend is clearly increasing, even if the existence of two periods characterised by several losses suggests to approach the results with some caution. Generally speaking, the Dixon and Coles model seems to show more fluctuations and a lower overall return: the latter is also evident by comparing the average percentage gain obtained by the Mar-Co model (2.92%) with that obtained by the Dixon and Coles model (0.63%).

4 Discussion

The aim of this work has been to develop a model that, moving from the Dixon and Coles approach, is capable of properly predicting outcomes of football matches. The introduction of a joint PMF that allows the number of goals scored by the two opposing teams to be dependent when the final results are different from 0-0, 1-0, 0-1, 1-1 is aimed at providing a more comprehensive representation of the phenomenon. The model suggests a flexible solution as the marginal PMFs of X and Y obtained from the mixture joint PMF are Poisson PMFs when $\theta_3 = 0$, whereas a dependence structure is included in the model when $\theta_3 \neq 0$.

On the other hand, different aspects related to the model are potential limits of the approach. First of all, fit of the mixture joint PMF and the marginal PMFs to football scores should be accurately investigated in order to evaluate the proposed assumptions. Secondly, an in-depth comparison between the two estimation methods discussed in Sect. 2 could help to determine the best one in terms of balance between formal and computational advantages.

In addition, several modifications may represent the focus of future works: for example, in this work we have considered only odds for (1-X-2), but nowadays bookmakers offer a wide range of alternatives (*e.g.* Asian Handicap odds) that allows to extend the set $E^{(t)}$. Also, the inclusion of further covariates in the model, such as information provided by bookmakers [2], should be considered in order to improve the overall predictive capability of the model.

However, the development of a betting strategy based on our model provides encouraging results, also if compared to those obtained by the Dixon and Coles model. The analysis here presented is surely not sufficient to state that an improvement of the *gold standard* model has been developed, but the apparent good results achieved represent a promising starting point for further analysis.

Acknowledgements The authors are grateful to David Dandolo for providing the code on the Dixon and Coles model and Nicola Sartori for advices.

References

1. Dixon, M.J., Coles, S.G.: Modelling association football scores and inefficiencies in the football betting market. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, **46**, 265–280 (1997)
2. Egidi, L., Pauli, F., Torelli, N.: Combining historical data and bookmakers' odds in modelling football scores. *Stat. Model.*, **18**(5–6), 436–459 (2018)
3. Gong, G., Samaniego, F.J.: Pseudo maximum likelihood estimation: theory and applications. *Ann. Statist.*, **9**(4), 861–869 (1981)
4. Karlis, D., Ntzoufras, I.: Analysis of sports data by using bivariate Poisson models. *J. R. Stat. Soc. Ser. D.*, **52**(3), 381–393 (2003)
5. Kelly, J.L.Jr.: A New Interpretation of Information Rate. *Bell Syst. Tech. J.*, **35**, 917–926 (1956)
6. Maher, M.J.: Modelling association football scores. *Stat. Neerl.*, **36**(3), 109–118 (1982)

Estimation of dynamic quantile models via the MM algorithm

Stima di modelli Quantilici Dinamici con algoritmo MM

Fabrizio Poggioni, Mauro Bernardi, Lea Petrella

Abstract Accurate Value-at-Risk measurement often requires estimation of complex dynamic models where usually the parameters enter nonlinearly the quantile estimation equation. In this paper we address the problem of estimation of the parameters of a class of conditionally autoregressive Value at risk models by adapting the Majorizing–Minorizing algorithm of Hunter and Lange (2000).

Abstract *Per ottenere delle valide misure di Value-at-Risk occorre stimare modelli dinamici complessi nei quali spesso il modello di regressione quantilica risulta non lineare nei parametri. In questo lavoro affrontiamo il problema di stima dei parametri per modelli di Value at risk condizionatamente autoregressivi, adattando l'algoritmo Majorizing–Minorizing proposto da Hunter and Lange (2000)*

Key words: Conditional autoregressive quantiles; CAViaR; MM algorithm; Value-at-Risk.

1 Introduction

Accurate risk measurement is a primary need for financial institutions and investors especially after the recent financial crisis. Within the instruments for market risk measurement, Value-at-Risk (VaR) is certainly one of the most popular and used approaches. For a given portfolio, probability and time horizon, VaR is defined as

Fabrizio Poggioni

Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome e-mail: fabrizio.poggioni@uniroma1.it

Mauro Bernardi

Department of Statistical Sciences, University of Padova

Lea Petrella

Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome

a threshold value such that the probability that the market-to-market loss on the portfolio over the given time horizon exceeds this value is the given probability level. Therefore, it can be simply understood as a specific conditional quantile of the portfolio returns given the current information, i.e., $\mathbb{P}(Y_t < -\text{VaR}_t \mid \mathcal{F}_t) = \tau$, where Y_t and \mathcal{F}_t denote respectively the return of a portfolio and the information set available at time t , while $\tau \in (0, 1)$ denotes the quantile confidence level associated with the VaR. There are several ways to estimate a VaR. The standard models initially developed to calculate the VaR of a portfolio can be categorized as: (i) the variance-covariance approach (Parametric method), (ii) Historical Simulation (Non-parametric method) and (iii) Monte Carlo simulation (semi-parametric method). Here, we focus on the CAViaR (Conditional Autoregressive Value at Risk) class of VaR models, introduced by Engle and Manganelli (2004). The inferential issue for the CAViaR class of models has been addressed in literature both from the frequentist and the Bayesian point of view. In the frequentist approach the CAViaR models allow to use the inferential Quantile Regression (QR) methods (Koenker 2005, Engle and Manganelli 2004) by minimizing the loss function introduced by Koenker and Basset (1978). A well-known problem related to the estimation of quantile regression models is the non full differentiability of the loss function, we propose in this regard an improvement in the estimation of CAViaR models via MM (Majorize-Minimization) algorithm, introduced by Ortega and Rheinboldt (1970) and adapted to the estimation of quantile regression models by Lange (2000).

2 Conditional Autoregressive Value-at-Risk models

Koenker (2005) showed how to embed the problem of estimating quantiles in a regression framework. Consider a sample of observations (y_1, y_2, \dots, y_T) generated by the model

$$y_t = \mathbf{x}_t^\top \beta + \varepsilon_t, \quad q_{t,\tau}(\varepsilon_t \mid \mathbf{x}_t) = 0, \quad (1)$$

where \mathbf{x}_t is a $p \times 1$ vector of regressors, β is a $p \times 1$ vector of regression coefficients, and $q_{t,\tau}(\varepsilon_t \mid \mathbf{x}_t)$ is the τ -quantile of ε_t conditional on \mathbf{x}_t . Let $f_t(\beta) \equiv \mathbf{x}_t^\top \beta$, then the τ -th regression quantile can be found by minimizing with respect to β the following :

$$\frac{1}{T} \sum_{t=1}^T \rho_\tau(Y_t - f_t(\beta)), \quad (2)$$

where $\rho_\tau(u) \equiv (\tau - \mathbb{1}_{(-\infty, 0)}(u))u$ is the well known check function. As emphasized in equation (2), the estimation method also works for non-linear specifications of $f_t(\beta)$, case that we deal with in this work. Let y_t be the return at time t , the CAViaR models introduced by Engle and Manganelli (2004) have the following form:

$$y_t = q_{t,\tau} + \varepsilon_t, \quad q_{t,\tau}(\varepsilon_t \mid \mathcal{F}_{t-1}) = 0 \quad (3)$$

$$q_{t,\tau} = \omega + \gamma q_{t-1,\tau} + \beta l(y_{t-1}), \quad (4)$$

where $q_{t,\tau}$ is the τ level quantile of y_t , (ω, γ, β) are regression parameters and $l(\cdot)$ is an unknown function of the past returns. Different CAViaR models have been proposed in literature that differ for the specification of the function $l(\cdot)$:

$$l(y_t) = \beta_1 |y_t| \quad \text{Symmetric Absolute Value} \quad (5)$$

$$l(y_t) = \beta_1 (y_t)^+ + \beta_2 (y_t)^- \quad \text{Asymmetric Slope} \quad (6)$$

$$l(y_t) = \begin{cases} \beta_1 |y_t|, & z_t \leq r \\ \beta_2 |y_t|, & z_t > r \end{cases} \quad \text{Threshold CAViaR.} \quad (7)$$

Estimation

Let $\{y_t, \mathbf{x}_t\}_{t=1}^T$ a sample of observations from the CAViaR process in equation (1), the estimated CAViaR parameters are the solution to the minimization problem:

$$\begin{aligned} \hat{\vartheta}_\tau &= \arg \min_{\vartheta_\tau} \frac{1}{T} \sum_{t=1}^T \rho_\tau(y_t - q_\tau(\mathbf{x}_t, \vartheta_\tau)) \\ &= \arg \min_{\vartheta_\tau} \frac{1}{T} \mathcal{V}_\tau(\vartheta_\tau). \end{aligned} \quad (8)$$

As previously noted the function $\rho_\tau(u)$ presents problems of derivability, in particular this function is not differentiable when $u = 0$. Hunter and Lange (2000), proposed to estimate the quantile regression parameters by applying the Majorising–Minimizing (MM) method. Suppose we want to minimise the objective function $\mathcal{L}(\vartheta) : \mathbb{R}^{p+1} \rightarrow \mathbb{R}$ and denote with $\vartheta^{(k)}$ the current iterate, the MM algorithm proceeds in two steps:

(i) create a surrogate (majorizer) function $\mathcal{G}(\vartheta | \vartheta^{(k)}) : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ that satisfies

$$(i.1) \quad \mathcal{G}(\vartheta^{(k)} | \vartheta^{(k)}) = \mathcal{L}(\vartheta^{(k)});$$

$$(i.2) \quad \mathcal{G}(\vartheta | \vartheta^{(k)}) \geq \mathcal{L}(\vartheta), \text{ for all } \vartheta;$$

$$(ii) \hat{\vartheta}^{(k+1)} = \arg \min_{\vartheta \in \mathbb{R}^{p+1}} \mathcal{G}(\vartheta | \vartheta^{(k)}).$$

Therefore, $\mathcal{G}(\vartheta^{(k+1)} | \vartheta^{(k)}) \leq \mathcal{G}(\vartheta^{(k)} | \vartheta^{(k)})$, with conditions (i) and (ii) entails the descent property, i.e., $\mathcal{L}(\vartheta^{(k+1)}) \leq \mathcal{L}(\vartheta^{(k)})$. For what about the quantile regression problem, the MM algorithm can be advantageous if a good approximation of $\mathcal{V}_\tau(\vartheta_\tau)$ as defined in equation (8) is found, in combination with a good majorizer function possibly fully derivable at each iteration. Hunter and Lange (2000) propose to approximate the quantile check function $\rho_\tau(u)$ which underlies the quantile loss function $\mathcal{V}_\tau(\vartheta_\tau)$, by its ε -perturbation, defined as

$$\rho_\tau^\varepsilon(u) = \rho_\tau(u) - \frac{\varepsilon}{2} \log(\varepsilon + |u|), \quad (9)$$

in such a way that $\mathcal{V}_\tau^\varepsilon(\vartheta_\tau) = \sum_{i=1}^n \rho_\tau^\varepsilon(\vartheta_\tau)$ approximate $\mathcal{V}_\tau(\vartheta_\tau)$ for any $\varepsilon > 0$. Let $\hat{\mathbf{r}}_\tau^{(k)} = \mathbf{y} - q_\tau(\mathbf{x}_t, \vartheta_\tau)$ the vector of quantile residual values at iteration k , then Hunter and Lange (2000) proposed to majorise the ε -perturbation of the quantile check function in equation (9) at $\hat{r}_{t,\tau}^{(k)}$ by the quadratic function

$$\zeta_\tau^\varepsilon(r_t | \hat{r}_{t,\tau}^{(k)}) = \frac{1}{4} \left[\frac{r_{t,\tau}^2}{\varepsilon + |\hat{r}_{t,\tau}^{(k)}|} + (4\tau - 2)r_{t,\tau} + c_\tau \right], \quad (10)$$

where $r_t = y_t - q_\tau(\mathbf{x}_t, \vartheta_\tau)$ and $\hat{r}_{t,\tau}^{(k)}$ is the i -th value of $\hat{\mathbf{r}}_\tau^{(k)}$ and the constant c_τ is chosen so that $\sum_{i=1}^n \zeta_\tau^\varepsilon(\hat{r}_{t,\tau}^{(k)} | \hat{r}_{t,\tau}^{(k)}) = \sum_{i=1}^n \rho_\tau^\varepsilon(\hat{r}_{t,\tau}^{(k)})$. The MM algorithm for quantile regression operates by minimising the majoriser

$$\begin{aligned} \hat{\vartheta}_\tau^{(k+1)} &= \arg \min_{\vartheta_\tau} \mathcal{G}_\tau^\varepsilon(\vartheta_\tau | \hat{\vartheta}_\tau^{(k)}), \\ \mathcal{G}_\tau^\varepsilon(\vartheta_\tau | \hat{\vartheta}_\tau^{(k)}) &= \frac{1}{n} \sum_{i=1}^n \zeta_\tau^\varepsilon(r_t | \hat{r}_{t,\tau}^{(k)}), \end{aligned} \quad (11)$$

with respect to ϑ_τ and the minimiser becomes the next iterate. The main property of the MM algorithm is that it delivers a sequence of parameters $\hat{\vartheta}_\tau^{(k)}$ for $k = 0, 1, 2, \dots$ that does not increase the loss function, i.e., $\ell(\hat{\vartheta}_\tau^{(k)}) \geq \ell(\hat{\vartheta}_\tau^{(k+1)})$. This property is satisfied even if the minimisation step in equation (11) is substituted by a step that reduces the loss function, leading to a Generalised MM (GMM) algorithm:

$$\hat{\vartheta}_\tau^{(k+1)} \ni \mathcal{G}_\tau^\varepsilon(\hat{\vartheta}_\tau^{(k+1)} | \hat{\vartheta}_\tau^{(k)}) \leq \mathcal{G}_\tau^\varepsilon(\hat{\vartheta}_\tau^{(k+1)} | \hat{\vartheta}_\tau^{(k)}). \quad (12)$$

The previous result becomes relevant for CAViaR estimation since the CAViaR dynamics is not linear in the parameters, therefore we need to resort to a Newton–Raphson update that exploits the gradient and hessian matrix. Specifically, let $\frac{\partial r_t}{\partial \vartheta_\tau} =$

$$-\frac{\partial q_\tau(\mathbf{x}_t, \vartheta)}{\partial \vartheta} \Big|_{\vartheta_\tau = \hat{\vartheta}_\tau} \text{ and } \frac{\partial^2 r_t}{\partial \vartheta_\tau \partial \vartheta_\tau} = \frac{\partial q_\tau(\mathbf{x}_t, \vartheta)}{\partial \vartheta} \frac{\partial q_\tau(\mathbf{x}_t, \vartheta)}{\partial \vartheta'} \Big|_{\vartheta_\tau = \hat{\vartheta}_\tau}, \text{ then}$$

$$\hat{\vartheta}_\tau^{(k+1)} = \hat{\vartheta}_\tau^{(k)} - \left[\mathcal{G}_\tau^{\tau''}(\vartheta_\tau | \hat{\vartheta}_\tau^{(k)}) \Big|_{\vartheta_\tau = \hat{\vartheta}_\tau^{(k)}} \right]^{-1} \mathcal{G}_\tau^{\tau'}(\vartheta_\tau | \hat{\vartheta}_\tau^{(k)}) \Big|_{\vartheta_\tau = \hat{\vartheta}_\tau^{(k)}}, \quad (13)$$

where $\mathcal{G}_\tau^{\tau'}(\vartheta_\tau | \hat{\vartheta}_\tau^{(k)}) \Big|_{\vartheta_\tau = \hat{\vartheta}_\tau^{(k)}}$, $\mathcal{G}_\tau^{\tau''}(\vartheta_\tau | \hat{\vartheta}_\tau^{(k)}) \Big|_{\vartheta_\tau = \hat{\vartheta}_\tau^{(k)}}$ are the gradient and hessian matrix, respectively and $\hat{r}_{t,\tau}^{(k)} = y_t - q_{t,\tau}(\mathbf{x}_t, \hat{\vartheta}_\tau^{(k)})$.

	<i>Symmetric</i>	<i>Asymmetric</i>	<i>Threshold</i>
$DQ_{\tau=0.05}$	0.4375	0.7423	0.5461
$DQ_{\tau=0.01}$	0.9818	0.9863	0.9679

Table 1 The Table shows the p-values related to the Dynamic Quantile test as introduced by Engle and Manganelli (2004). The first row is relative to the quantile level $\tau = 0.05$, the second row to the quantile level $\tau = 0.01$.

3 Empirical Application

Here we present an empirical application of the MM-estimation methodology for CAViaR models to daily financial returns. We considered the Exxon Mobil Corporation (XOM), an institutions belonging to the S&P500 composite index. We perform estimation and back-testing of the 3 presented CAViaR model specifications for two quantile confidence levels $\tau = (0.01, 0.05)$. We considered 6482 daily observations for the time lapse

5/10/1993 - 08/08/2018. The first 4000 observation are employed to estimate the regression parameters, the last 2498 are employed to evaluate the out-of-sample performances. Parameter estimation are performed via the MM algorithm as introduced in the previous sections. In Fig. 1 we show the 3 out-of-the results, i.e. the forecasts of each of the 3 considered models. In Table 1 we report the Dynamic

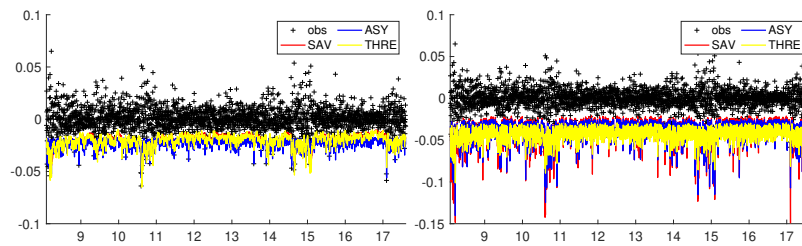


Fig. 1 The forecasts of the conditional quantiles according to the three considered models. In the left quadrant with respect to the quantile level $\tau = 0.05$, in the left quadrant with respect to the quantile level $\tau = 0.01$

Quantile test results as introduced by Engle and Manganelli (2004), looking at the p-values in the table we can see how the proposed models present good forecast performances for both the considered quantile levels.

4 Conclusion

CAVIAR models have been successful in financial applications. Following the introduction of these models, many authors have proposed different specifications of

these dynamic models for the prediction of Value-at-Risk for financial data. We proposed to apply the MM algorithm to solve some known estimation problems of quantile regressions, focusing on some dynamic Quantile Regression models specifically dedicated to financial data. Up to our knowledge this is the first attempt in the literature to connect the MM algorithm to the quantile models with dynamic structure. We aim to extend this methodology to the entire class of CAVIAR models present in the literature.

References

1. Engle and Manganelli: Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22:367–381 (2004).
2. Engle and Ng: Measuring and testing the impact of news on volatility. *The Journal of Finance*, 48(5):1749–1778 (1993).
3. Hunter and Lange: Quantile regression via an MM algorithm. *J. Comput. Graph. Statist*, 9(1):60–77 (2000).
4. Hunter and Lange: A tutorial on MM algorithms. *Amer. Statist.*, 58(1):30–37 (2004).
5. Koenker: *Quantile Regression*. Cambridge University Press, Cambridge.
6. Koenker and Basset: Regression quantiles. *Econometrica*, 46:33–50 (1978).
7. Lange, Kenneth , Wu, Tong Tong and others: The MM alternative to EM. *Statistical Science*, 492–505 (2010).
8. Ortega and Rheinboldt: Iterative solution of nonlinear equations in several variables. volume 30 *Siam* (1970).

The decomposition by subpopulations of the Pietra index: an application to the professional football teams in Italy

La scomposizione per sottopopolazioni dell'indice di Pietra: un'applicazione alle squadre professionistiche di calcio in Italia

Francesco Porro and Mariangela Zenga

Abstract

In this paper an innovative method to decompose by subpopulations the Pietra index is presented. The decomposition procedure is based on a two-step procedure that has been used in the literature to decompose also other inequality indexes. A such decomposition allows the evaluation of the contributions related to each subpopulation to the total value of the index. Beyond the methodological details, an application with real data about the professional football teams in Italy is described.

Abstract

Questo articolo presenta un innovativo metodo di scomposizione per sottopopolazioni dell'indice di Pietra. La scomposizione si basa su una procedura a due fasi già utilizzata in letteratura per scomporre altri indici di disuguaglianza. Tale scomposizione permette la valutazione dei contributi al valore totale dell'indice, associati ad ogni sottopopolazione. Oltre ai dettagli metodologici, viene descritta anche un'applicazione riguardante le squadre professionistiche di calcio in Italia.

Key words: Pietra index, Decomposition by subpopulations

1 Introduction

The Pietra index is a very well-known index. Originally, it was proposed in the literature in [3] to evaluate the inequality in the field of income distribution analysis, but it has been used in a very large number of applications. Its simple interpretation

Francesco Porro

Università degli Studi di Milano-Bicocca, piazza dell'Ateneo Nuovo 1, Milano e-mail: francesco.porro1@unimib.it

Mariangela Zenga

Università degli Studi di Milano-Bicocca, piazza dell'Ateneo Nuovo 1, Milano e-mail: mariangela.zenga@unimib.it

is surely a reason for its longevity. The value of the Pietra index is the proportion of the total amount T that if it is transferred (and properly redistributed) from the values greater than the mean to the other values, it bears to the egalitarian situation, where all the units have the same amount. As for all the inequality indexes, the decomposition by subpopulations of the Pietra index is a very useful tool. In the literature there are some decompositions of this index (see for example [1], and [2]), but the procedure presented in this paper can be considered easier and more versatile.

2 Notation and initial settings

Let Y be a non-negative statistical variable on a population of size N . Let $y_1 < y_2 < \dots < y_r$ denote the distinct values assumed by Y with frequencies n_1, n_2, \dots, n_r , respectively. Obviously, $\sum_{h=1}^r n_h = N$. Let $M(Y) = \sum_{h=1}^r y_h \cdot \frac{n_h}{N}$ be the arithmetic mean of Y and let

$$S_{M(Y)} = \sum_{h=1}^r |y_h - M(Y)| \cdot \frac{n_h}{N}$$

be the mean absolute deviation of Y from $M(Y)$. The inequality index \mathcal{P} proposed by Pietra in [3] is defined as:

$$\mathcal{P} = \frac{S_{M(Y)}}{2M(Y)}.$$

Now, since

$$\sum_{h=1}^r |y_h - M(Y)| n_h = \sum_{\{y_h : y_h \leq M(Y)\}} (M(Y) - y_h) n_h + \sum_{\{y_h : y_h > M(Y)\}} (y_h - M(Y)) n_h$$

and

$$\sum_{\{y_h > M(Y)\}} (y_h - M(Y)) n_h = - \sum_{\{y_h \leq M(Y)\}} (y_h - M(Y)) n_h,$$

it holds also that

$$\sum_{h=1}^r |y_h - M(Y)| n_h = 2 \cdot \sum_{\{y_h : y_h \leq M(Y)\}} (M(Y) - y_h) n_h.$$

For this reason, it is easy to see that the Pietra index can also be calculated as:

$$\mathcal{P} = \frac{1}{\sum_{h=1}^r y_h n_h} \cdot \sum_{\{y_h : y_h \leq M(Y)\}} (M(Y) - y_h) n_h. \quad (1)$$

At each y_h (with $h = 1, 2, \dots, r$) the whole population can be splitted in two non-overlapping groups, the lower group and the upper one:

- the first one, corresponding to $\{Y \leq y_h\}$, containing the first $P_h = \sum_{t=1}^h n_t$ units;

The decomposition by subgroups of the Pietra index

- the second one, corresponding to $\{Y > y_h\}$, containing the other $N - P_h$ units.

If now we denote $\bar{h} = \max\{h : y_h \leq M(Y)\}$, then the formula (1) becomes:

$$\mathcal{P} = \frac{M(Y) - \bar{M}_{\bar{h}}(Y)}{M(Y)} \cdot P_{\bar{h}}, \quad (2)$$

$$\text{where: } P_{\bar{h}} = \frac{P_{\bar{h}}}{N} = \frac{\sum_{h=1}^{\bar{h}} n_h}{N} \quad \text{and} \quad \bar{M}_{\bar{h}}(Y) = \frac{\sum_{h=1}^{\bar{h}} y_h n_h}{P_{\bar{h}}} \quad (3)$$

are the cumulative relative frequency corresponding to $y_{\bar{h}}$, and the lower mean at \bar{h} (that is the arithmetic mean of lower group $\{Y \leq M(Y)\}$), respectively. It is interesting to remark that $\frac{M(Y) - \bar{M}_{\bar{h}}(Y)}{M(Y)}$ in formula (2) is the pointwise inequality measure of the Bonferroni index, at $h = \bar{h}$. More details on this point can be found in [4].

3 The decomposition by subpopulations of the Pietra index

In this section the procedure to decompose the Pietra index by subpopulations is described, starting from the formula (2). By using the same setting of the previous section, we consider now a population with N units, divided into k (with $k \geq 2$) different subpopulations. The value r is the number of distinct values of the variable Y , and n_{hg} denotes the frequency of the value y_h in the subpopulation g , (obviously $h = 1, \dots, r$ and $g = 1, \dots, k$). For the distribution of the subpopulation g , we can

Table 1 Bivariate distribution of the variable Y , according to the k subpopulations

Y	Subpopulations				
	1	2	...	k	
y_1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
y_2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
\vdots	\vdots	\vdots	...	\vdots	\vdots
y_r	n_{r1}	n_{r2}	...	n_{rk}	$n_{r.}$
	$n_{.1}$	$n_{.2}$...	$n_{.k}$	N

consider the quantities $P_{hg} = \sum_{t=1}^h n_{tg}$, and $Q_{hg} = \sum_{t=1}^h y_t n_{tg}$, that are the cumulative frequency of y_h in the subpopulation g , and the sum of the values of the lower group $\{Y \leq y_h\}$ in the subpopulation g . It follows that we can define the lower mean $\bar{M}_{hg}(Y)$ for the subpopulation g as:

$$\bar{M}_{hg}(Y) = \begin{cases} 0 & \text{if } P_{hg} = 0 \\ \frac{Q_{hg}(Y)}{P_{hg}} & \text{if } P_{hg} > 0 \end{cases}$$

and the two ratios $\frac{n_{\cdot g}}{N} = \frac{\sum_{h=1}^r n_{hg}}{N}$ and $p(g|h) = \frac{P_{hg}}{P_{h\cdot}}$ are the relative frequencies of the subpopulation g in the whole population and in the lower group $\{Y \leq y_h\}$, respectively. The following relationship among the mean $M(Y)$ and the means of the k subpopulations $M_g(Y)$ holds true

$$M(Y) = \sum_{g=1}^k M_g(Y) \cdot \frac{n_{\cdot g}}{N}, \quad (4)$$

while among the lower mean $\bar{M}_{\bar{h}\cdot}(Y)$ and the k lower means $\bar{M}_{\bar{h}l}(Y)$, it holds that:

$$\bar{M}_{\bar{h}\cdot}(Y) = \sum_{l=1}^k \bar{M}_{\bar{h}l}(Y) \cdot p(l|\bar{h}).$$

By using this last relationship in the formula (2) of the Pietra index, and by recalling that $\sum_{l=1}^k p(l|\bar{h}) = 1$, we have:

$$\begin{aligned} \mathcal{P} &= \frac{M(Y) - \bar{M}_{\bar{h}\cdot}(Y)}{M(Y)} \cdot p_{\bar{h}} = \frac{\sum_{l=1}^k [M(Y)p(l|\bar{h}) - \bar{M}_{\bar{h}l}(Y)p(l|\bar{h})]}{M(Y)} \cdot p_{\bar{h}} \\ &= \sum_{l=1}^k \frac{M(Y) - \bar{M}_{\bar{h}l}(Y)}{M(Y)} \cdot p(l|\bar{h}) \cdot p_{\bar{h}} = \sum_{l=1}^k V_{\bar{h}l}(Y) \cdot p_{\bar{h}}, \quad \text{where} \\ V_{\bar{h}l}(Y) \cdot p_{\bar{h}} &= \frac{M(Y) - \bar{M}_{\bar{h}l}(Y)}{M(Y)} \cdot p(l|\bar{h}) \cdot p_{\bar{h}} \end{aligned} \quad (5)$$

can be interpreted as the contribution of the subpopulation l to the Pietra index. Using this decomposition procedure is therefore possible to assess the contribution to the total value of \mathcal{P} related to each single subpopulation, provided by formula (5). Such result is important, because the most of the decomposition methods in the literature are not able to reach this goal. Now, recalling the formula (4), and that $\sum_{g=1}^k \frac{n_{\cdot g}}{N} = 1$, we have that:

$$\mathcal{P} = \sum_{l=1}^k \sum_{g=1}^k \left[\frac{M_g(Y) - \bar{M}_{\bar{h}l}(Y)}{M(Y)} \cdot \frac{n_{\cdot g}}{N} \cdot p(l|\bar{h}) \right] \cdot p_{\bar{h}} = \sum_{l=1}^k \sum_{g=1}^k V_{\bar{h}lg}(Y) \cdot p_{\bar{h}}. \quad (6)$$

By using formula (6), the decomposition of the Pietra index as $\mathcal{P} = \mathcal{P}_W + \mathcal{P}_B$ can be obtained, where:

$$\mathcal{P}_W = \sum_{l=1}^k V_{\bar{h}ll}(Y) \cdot p_{\bar{h}} = \sum_{l=1}^k \frac{M_l(Y) - \bar{M}_{\bar{h}l}(Y)}{M(Y)} \cdot \frac{n_{\cdot l}}{N} \cdot p(l|\bar{h}) \cdot p_{\bar{h}}$$

The decomposition by subgroups of the Pietra index

$$\mathcal{P}_B = \sum_{l=1}^k \sum_{g \neq l} V_{hlg}(Y) \cdot p_{\bar{h}} = \sum_{l=1}^k \sum_{g \neq l} \frac{M_g(Y) - \bar{M}_{hl}(Y)}{M(Y)} \cdot \frac{n_g}{N} \cdot p(l|\bar{h}) \cdot p_{\bar{h}}$$

are the Within and Between part of the Pietra index, respectively, since in the first one (\mathcal{P}_W), only the means of the same subpopulation are compared, while in the second one (\mathcal{P}_B), the comparisons are between means of different subpopulations.

4 An application to the professional football teams in Italy

The decomposition by subpopulations presented in the previous sections is here applied to the Italian professional football teams. We consider the 20 teams in the major league (named "Serie A"), the 19 teams in the second one ("Serie B") and finally the 59 teams in the third one ("Serie C"). Actually this last league is territorially divided into three more subgroups, but in this application, we examine all the teams in these three groups together. For each team we consider the variable Y = "value (in euro) of the players of the team", which is the sum of the values of all the team players. Data are available in the website: *www.transfermarkt.it*. In this application, we consider as subpopulations the three leagues: "Serie A", "Serie B" and "Serie C". In Table 2 some statistical indicators for the variable Y regarding the three subpopulations and the whole population are provided. From the calculations it arises that $\bar{h} = \max\{h : y_h \leq M(Y)\} = 80$. The values of the lower means $\bar{M}_{hl}(Y)$ and of $p(l|\bar{h})$ that we need for the calculations, are also reported in Table 2. From

Table 2 Summary of the variable Y in the subgroups and in the whole population

	Serie A ($l = 1$)	Serie B ($l = 2$)	Serie C ($l = 3$)	TOTAL
Mean	$M_1(Y) = 234.986$	$M_2(Y) = 17.556$	$M_3(Y) = 4.017$	$M(Y) = 53.778$
Size	$n_1 = 20$	$n_2 = 19$	$n_3 = 59$	$N = 98$
Median	$Me_1 = 146.765$	$Me_2 = 15.85$	$Me_3 = 3.8$	$Me(Y) = 5.155$
Mean Square Error	214.221	7.165	1.375	132.412
$\bar{M}_{80,l}$	40.825	17.556	4.017	
$p(l 80)$	0.025	0.2375	0.7375	

an overview of the data, it is worth noting that the overlappings in these three subpopulation are only two, since all the values of teams in Serie A are greater than the values of all the teams in the other two subpopulations, and there is only one team in Serie C, with value greater than the poorest team in Serie B. From direct computation we find that for the whole population of football professional teams, $S_{M(Y)} = 74.408$, therefore the Pietra index is:

$$\mathcal{P} = \frac{74.408}{2 \cdot 53.778} = 0.6918.$$

By applying the decomposition procedure presented in the previous sections, we obtain the decomposition of the Pietra index as $\mathcal{P} = \sum_{l=1}^3 \sum_{g=1}^3 V_{hlg}(Y) \cdot p_{\bar{h}}$, with the values of $V_{hlg}(Y) \cdot p_{\bar{h}}$ reported in Table 3. It is worth noting that the values $V_{h22}(Y)$ and $V_{h33}(Y)$ are zero, since the means and the lower means occurring in the calculation of these two quantities are equals. The aggregated values in the last row show that the subpopulation with the biggest contribution to the Pietra index is Serie C (since $V_{h3.}(Y) \cdot p_{\bar{h}} = 0.5564$), and the one with the smallest contribution is Serie A (as $V_{h1.}(Y) \cdot p_{\bar{h}} = 0.0049$). As special case, also the decomposition in the Within and Between parts can be obtained. The former is the sum of the entries in the main diagonal of Table 3, the latter the sum of all the remaining ones:

$$\mathcal{P}_W = 0.0150 \quad \text{and} \quad \mathcal{P}_B = 0.6768.$$

From these values, it is interesting to evaluate the two ratios

$$\frac{\mathcal{P}_W}{\mathcal{P}} = \frac{0.0150}{0.6918} = 0.022 \quad \text{and} \quad \frac{\mathcal{P}_B}{\mathcal{P}} = \frac{0.6768}{0.6918} = 0.978,$$

which show that in this application the Between part of the Pietra index is largely more relevant than the Within one, since it represents the 97.8% of the total index, while the Within part only the 2.2%.

Table 3 The values of the quantities $V_{80lg}(Y) \cdot p_{80}$

	1	1	2	3
g				
1		0.0150	0.1597	0.5271
2		-0.0017	0	0.0293
3		-0.0084	-0.0292	0
$V_{80l.}(Y) \cdot p_{80}$		0.0049	0.1305	0.5564

References

1. Arnold, B. C., Sarabia, J.M.: Majorization and the Lorenz order with applications in applied mathematics and economics. Cham, Switzerland, Springer (2018)
2. Frosini, B.V.: Approximation and decomposition of Gini, Pietra–Ricci and Theil inequality measures, Empirical Economics. **43**, 175–197 (2012)
3. Pietra, G.: Delle relazioni fra indici di variabilit , note I e II. Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti. **74**, 775–804 (1915)
4. Valli, I., Zenga, M.: Joint decomposition by subpopulations and sources of the point and synthetic Bonferroni inequality measures, Statistica & Applicazioni. **15(2)**, 83–120 (2017)

An Object Oriented Data Analysis of Tweets: the Case of Queen Elizabeth Olympic Park

Object Oriented Data Analysis di Tweet: il caso del Queen Elizabeth Olympic Park

Paola Riva, Paola Sturla, Anna Calissano and Simone Vantini

Abstract The talk will focus on the problem of embedding human language complexities in the framework of Object Oriented Data Analysis, showing a novel application of these statistical techniques to textual data. Starting from the investigation of the vast literature dealing with natural language understanding, we will define a taxonomy of mathematical models aimed at embedding complex properties of text. Along the talk, we will also discuss the potentials of the statistical analysis of the content of social media messages to support the landscape design process, showing the results of the analysis of Tweets collected from the Queen Elizabeth Olympic Park in London. This contributes to the conversation on the employment of Big Data analytical techniques to study the users perception of landscape.

Abstract *La presentazione sarà focalizzata sul problema di contestualizzare le complessità del linguaggio umano all'interno dell'Object Oriented Data Analysis, mostrando un'applicazione innovativa di queste tecniche al caso di dati testuali. A partire dall'esplorazione della vasta letteratura interessata alla comprensione del linguaggio umano, definiremo una tassonomia di modelli matematici mirati ad inglobare proprietà complesse del testo. Durante la presentazione discuteremo anche le potenzialità di svolgere un'analisi statistica del contenuto di messaggi derivanti da social media per supportare il processo di progettazione del paesaggio urbano,*

Paola Riva

DIG - Dept. of Management, Economics and Industrial Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133 MI, Italy, e-mail: paola.riva@polimi.it

Paola Sturla

Dept. Landscape Architecture, Harvard Graduate School of Design, 48 Quincy Street, Gund Hall 422, Cambridge, MA 02138, USA, e-mail: psturla@gsd.harvard.edu

Anna Calissano

MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133 MI, Italy, e-mail: anna.calissano@polimi.it

Simone Vantini

MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133 MI, Italy, e-mail: simone.vantini@polimi.it

mostrando i risultati dell'analisi di Tweet derivanti dal Queen Elizabeth Olympic Park di Londra. Ciò contribuisce al dibattito sull'utilizzo di tecniche di analisi di Big Data per lo studio della percezione del paesaggio da parte del visitatore.

Key words: Object Oriented Data Analysis, Statistical Text Analysis, Text Embedding, Digital Landscape Architecture, Landscape Perception, Queen Elizabeth Olympic Park

1 Motivation

Within the context of *data complexity*, Object Oriented Data Analysis (OODA) offers an innovative approach to deal with populations of complex objects, whose statistical analysis require the definition of mathematical spaces with enhanced properties with respect to the one of a standard Euclidean space [26, 13, 14]. In recent years, OODA has been employed to investigate a variety of complex statistical units, such as functions [21, 22, 7, 24, 15], images [10, 27], trees [25, 26], shapes [6, 4, 17, 28], covariance operators [5] and data lying in manifolds [8, 11, 12]. Along this talk, we will present the study we conducted to approach the analysis of *text* from an OODA perspective. In particular, we will discuss the complicated issue of embedding complex properties of human language, such as syntax and semantics, into complex statistical units to be analysed in the setting of OODA. Therefore, this study offers a novel approach to the statistical analysis of textual data, posing the attention on the importance of embedding complex properties of text in the statistical atom to be further analysed in the setting of OODA.

2 Talk Outline

The first part of the talk will focus on the problem of framing complex properties of human language, such as syntax and semantics, in the context of Object Oriented Data Analysis (OODA). This will open a discussion on the question of representing complexities of text in suitable mathematical spaces, offering an opportunity to present an overview of extant literature about text embedding. In particular, we will individuate mathematical models aimed at embedding the meaning of text (i.e. semantic models) and models concerned with the representation of the structure of sentences (i.e. syntactic models). Thanks to this investigation, we will be able to present a taxonomy of mathematical models for embedding either semantics or syntax of text.

Besides presenting an overview of existing models for embedding text properties, this work contributes to the conversation over the digitalization of Landscape Architecture [20, 18, 19, 1, 2], taking part to the discussion on the potential use of Big Data techniques in supporting landscape studies [23, 9, 16]. Indeed, this talk will

discuss the potential of enriching the study of users' landscape perception applying OODA methodologies to analyze the content of social media messages generated by actual visitors of a specific area. In particular, we will present the results of a statistical analysis of geo-localised Tweets collected along three weeks of 2018 from the Queen Elizabeth Olympic Park (QEOP) in London. Starting from the description of the procedure employed to gather geo-localised data from Twitter, we will discuss the limitations of this data collection technique and the importance of selecting a data sample that is rich in information relevant for the application, even if this may imply significant reductions in the size of dataset. We will also show the role played by the mathematical model to embed the semantic of text in exploiting the information contained in the message of Tweets. This will open a discussion on the importance of selecting a proper set of variables for capturing the characteristic features to represent data complexity. Finally, we will present the results of a non-hierarchical cluster analysis (k-medoids) applied to the text of Tweets, whose meaning has been embedded using Latent Semantic Analysis [3]. Without assuming any prior difference among Tweets, we find the existence of four clusters of semantically similar messages. In particular, we observed that two groups of Twitter users in the selected area are interested in events related to the London Stadium (i.e., football and rugby matches), one is mainly concerned with shopping due to the presence of an important city mall close to the park (i.e. Westfield Stratford City Mall) and a group talks about the Queen Elizabeth Olympic Park. The discussion of these results will highlight the potential contribution of this work to the conversation on the use of Big Data in Landscape Architecture.

References

- [1] Batty M (2007) *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. The MIT press
- [2] Batty M (2013) *The new science of cities*. Mit Press
- [3] Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407, DOI 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASII>3.0.CO;2-9
- [4] Dryden I, Mardia K (2016) *Statistical shape analysis, with applications in R: Second edition*. DOI 10.1002/9781119072492
- [5] Dryden I, Koloydenko A, Zhou D (2009) Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Annals of Applied Statistics* 3(3):1102–1123, DOI 10.1214/09-AOAS249
- [6] Dryden IL, Mardia KV (1998) *Statistical shape analysis, vol 4*. Wiley Chichester
- [7] Ferraty F, Vieu P (2006) *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media

- [8] Fletcher PT, Lu C, Pizer SM, Joshi S (2004) Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging* 23(8):995–1005
- [9] Kaußen L (2018) Landscape perception and construction in social media: An analysis of user-generated content. *Journal of Digital Landscape Architecture* 3:373–379, DOI 10.14627/537642040
- [10] Locantore N, Marron J, Simpson D, Tripoli N, Zhang J, Cohen K, Boente G, Fraiman R, Brumback B, Croux C, et al (1999) Robust principal component analysis for functional data. *Test* 8(1):1–73
- [11] Mardia KV (1975) Statistics of directional data. *Journal of the Royal Statistical Society Series B (Methodological)* pp 349–393
- [12] Mardia KV, Jupp PE (2009) *Directional statistics*, vol 494. John Wiley & Sons
- [13] Marron J, Alonso A (2014) Overview of object oriented data analysis. *Biometrical Journal* 56(5):732–753, DOI 10.1002/bimj.201300072
- [14] Marron J, Alonso A (2014) Rejoinder to the discussion of: Overview of object-oriented data analysis. *Biometrical Journal* 56(5):790–791, DOI 10.1002/bimj.201400113
- [15] Menafoglio A, Secchi P (2017) Statistical analysis of complex and spatially dependent data: A review of object oriented spatial statistics. *European journal of operational research* 258(2):401–410
- [16] Montaña F (2018) The use of geo-located photos as a source to assess the landscape perception of locals and tourists case studies: Two public open spaces in munich, germany. *Journal of Digital Landscape Architecture* 3:346–355, DOI 10.14627/537642037
- [17] Pizer SM, Thall AL, Chen DT (2000) M-reps: A new object representation for graphics. In: *ACM Transactions on Graphics*, Citeseer
- [18] Portugali J (2000) Spatial cognitive dissonance and sociospatial emergence in a self-organizing city. In: *Self-Organization and the City*, Springer, pp 141–173
- [19] Portugali J (2011) *Complexity, cognition and the city*. Springer Science & Business Media
- [20] Portugali J, Meyer H, Stolk E, Tan E (2012) *Complexity theories of cities have come of age: an overview with implications to urban planning and design*. Springer Science & Business Media
- [21] Ramsay JO, Silverman BW (2002) *Applied Functional Data Analysis: Methods and Case Studies*. Springer
- [22] Ramsay JO, Silverman BW (2005) *Functional Data Analysis*. Springer
- [23] Roth M, Hildebrandt S, Röhner S, Tilk C, von Raumer HGS, Roser F, Borsdorff M (2018) Landscape as an area as perceived by people: Empirically-based nationwide modelling of scenic landscape quality in germany. *Journal of Digital Landscape Architecture* 3:129–137, DOI 10.14627/537642014
- [24] Sangalli L, Secchi P, Vantini S, Veneziani A (2009) A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *Journal of the American Statistical Association* 104(485):37–48, DOI 10.1198/jasa.2009.0002

- [25] Wang H (2003) Functional data analysis of populations of tree-structured objects. PhD thesis, University of North Carolina at Chapel Hill
- [26] Wang H, Marron JS (2007) Object oriented data analysis: Sets of trees. *Ann Statist* 35(5):1849–1873, DOI 10.1214/009053607000000217
- [27] Wei S, Lee C, Wichers L, Marron J (2016) Direction-projection-permutation for high-dimensional hypothesis tests. *Journal of Computational and Graphical Statistics* 25(2):549–569, DOI 10.1080/10618600.2015.1027773
- [28] Yushkevich P, Pizer SM, Joshi S, Marron JS (2001) Intuitive, localized analysis of shape variability. In: *Biennial International Conference on Information Processing in Medical Imaging*, Springer, pp 402–408

Bias reduced estimation of a fixed effects model for Expected Goals in association football

Stima non distorta di un modello Expected Goal con effetti fissi nel calcio

Lorenzo Schiavon and Nicola Sartori

Abstract In recent years, there has been an increasing interest in the scientific evaluation of association football. This paper addresses the issue of assessing the scoring probabilities of shots, commonly named as Expected Goals, by specifying a suitable generalized linear model. In particular, in order to account for different players abilities, a fixed effects model is considered. This has the advantage of avoiding the potential bias induced by endogeneity in a random effects specification. On the other hand, it is well known that in fixed effects models standard likelihood inference is not reliable, due to the large dimension of the parameter. Therefore, we propose to estimate the model using Firth's bias reduction method. Recent theoretical results guarantee that bias reduced estimates are accurate even in this extreme scenario.

Abstract Recentemente si sta manifestando un crescente interesse nella valutazione scientifica del calcio. In questo articolo, si propone un modello lineare generalizzato per descrivere le probabilità di realizzazione dei tiri o *Expected Goals*. Un modello ad effetti fissi permette di quantificare le diverse abilità dei vari giocatori, evitando la possibile endogeneità conseguente ad una specificazione ad effetti casuali. D'altra parte, l'introduzione di effetti fissi comporta un'elevata dimensionalità dei parametri che rende inaffidabile l'utilizzo dell'usuale inferenza basata sulla verosimiglianza. Perciò, si propone di stimare il modello tramite il metodo di riduzione della distorsione proposto da Firth. Recenti risultati teorici garantiscono, anche in queste scenario estremo, le buone proprietà delle stime ottenute.

Key words: adjusted score equations, bias reduction, complementary log-log link, football (soccer), generalized linear model, shots

Lorenzo Schiavon

Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy, e-mail: lorenzo.schiavon@phd.unipd.it

Nicola Sartori

Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy, e-mail: sartori@stat.unipd.it

1 Introduction

Expected Goals (XGs) may be defined as statistical metrics that quantify the probability to convert shots in goals in association football matches. This measure has been developed both in academia (see for instance [12]) and by football analysts (as in [1]). Generally, XGs are based on a logistic regression model where the outcome variable represents the outcome of the shot (1 for goal, 0 otherwise), whereas the explanatory variables are the shot features.

One of the critical issues of these models lies in the heterogeneity among different players. Thus, a possible solution might be to consider fixed effects for the single players when the data have repeated measurements for several players. Such specification entails a large number of incidental parameters, which causes biased estimates, as it is known since Neyman and Scott [11]. Therefore, this paper addresses the definition of an XGs model with fixed effects and proposes to estimate its parameter using a bias reduction method. A random effects specification (as in [9]) could represent an alternative. However, since a correlation between covariates and the individual random effects cannot be excluded, there would be a risk of endogeneity, which would lead to biased estimates (see [10] for a discussion).

The data used in this paper regard 379 matches in the Serie A league season 2016/2017 and are provided by Instat [4]. Each observation is determined by a single shot, according to Instat definition. After preprocessing, the dataset has $p = 18$ covariates and $n = 7716$ observations, grouped by $q = 216$ different players. The data are highly unbalanced, with the number of shots for every player ranging between 4 and 166. Shots transformed in goals are 960, about 12% of the total observations.

It is worth nothing that, if a complete knowledge of positions and speeds of players and balls at the moment of the shot was accessible, other information about the shot and the action preceding it would be pointless to estimate scoring probabilities. Unfortunately, in this dataset, as is often the case for semi-automatic recording methods, information about defenders' and goalkeeper's position is missing. Several studies (see for example [12, 15]) have identified these variables as very influential on scoring probabilities. In view of this, it is important to try to approximate the information of defenders' position, or at least of the defensive pressure, by an appropriate use of explanatory variables. Two different groups of covariates will be used. The first group considers features which represent information immediately related to shot characteristics, as for instance *angle* and *distance* of the shot from the goal. The second group of features is the set of covariates that describe the action before the shot, as *set play*, *dribbling* or the type of *assist*, if present, that anticipates the shot. They do not directly influence the scoring probability, but they are useful to approximate missing information. Moreover, in order to remedy to the lack of information about defensive pressure, an additional covariate reflecting the teams' behaviour has been added. This feature represents the propensity of the shot to be blocked, by using the frequencies of blocked shots by teams over the season.

The article is organized as follows: in Sect. 2 the model and the methodology are presented; Sect. 3 illustrates the application to the Instat data. Finally, Sect. 4 gives a discussion of the results and highlights the limitations of the proposed method.

2 Bias reduced estimation of fixed effects model

Previous work on XGs [12, 15] suggests the use of a generalized linear model for binary response. This specification is justified by the interest in modelling probabilities and not merely in classifying shots as goals/no-goals. Starting from the paper of McHale and Szczepański [9], it could be reasonable to stratify the observations by players in order to properly account for heterogeneity. The stratification is obtained by including fixed effects $\alpha = (\alpha_1, \dots, \alpha_q)$ in the linear predictor η of the generalized linear model, which is defined as follows:

$$g(\pi_{ij}) = \eta_{ij}, \quad \eta_{ij} = \alpha_j + x_{ij}^T \beta,$$

with $i = 1, 2, \dots, m_j, \quad j = 1, 2, \dots, q \quad \text{and} \quad n = \sum_{j=1}^q m_j$

where $g(\pi_{ij})$ is a transformation of the scoring probability π_{ij} for the i -th shot of the j -th player, x_{ij} is a p -dimensional vector of covariates and β is a parameter vector.

The inclusion of parameter α allows an interesting interpretation. Indeed, the parameter α_j can be seen as a measure of the individual ability to convert shots in goals for the j -th player, since his scoring probability increases with the value of α_j .

On the other hand, this specification implies that the number of parameters $q + p$ can be considerable with respect to the sample size n when q is large and the m_j are moderate. As a consequence, the usual Maximum Likelihood (ML) estimates may be severely biased (see [11] and [13]). One way to reduce such bias is to apply bias reduction methods (BR), as proposed by Firth [2].

The method is based on an adjusted score function, in which the model-based adjustment is constructed with the aim of reducing the leading term of the bias of the ML estimator. Lunardon [8] has recently proved the good properties of this procedure in stratified settings such as those considered here. Such properties are similar to those obtained using modification of the profile likelihood for β [13]. However, the advantage of the BR method is that it yields better estimates of the fixed effects α as well, which are also of interest in the present application. Kosmidis et al. [7] give a general application of the method to generalized linear models, of which the R package `brglm2` [6] provides an efficient implementation. Specifically, the model chosen for the application presented in this paper is the complementary log-log binary regression model, which is often suitable for data characterized by a relatively low frequency of success.

A simulation study with four covariates and 900 observations grouped in 30 unbalanced strata was performed to obtain an empirical confirmation of the theoretical properties of BR. The results, not reported here, confirm the theoretical findings and highlight how the method strongly reduce the bias of all parameters with respect to ML estimator.

3 Application and results

The methodology illustrated in Sect. 2 has been applied to the data described in Sect. 1, by defining the linear predictor η as a linear combination of fixed effects and polynomial functions of some explanatory variables.

Three main principles have been followed in the choice of the explanatory variables. Firstly, some classification models, which automatically select features (e.g. random forest and polymars [3]), have been estimated to determine the most important variables and their interactions. Then, stepwise procedures have been implemented to optimize information criteria or predictive scores on a test set. Lastly, also a qualitative evaluation, based on football knowledge, has been considered to ensure that the chosen covariates allow to interpret the associated parameters. The set of variable selected includes polynomial functions of *distance* and *angle* of the shot, type of shot and action, length and angle of *pass*, the chance that the shot is blocked and the interaction between *distance* and type of *pass*.

In order to evaluate the accuracy of the method, predictive scores have been compared on a subset (test set) of data. It should be noted that this type of comparison is particularly important in classification problems, but it is not necessarily sensible in the case of this paper, where forecasting appears pointless. Indeed, it is preferable to assess correctly probabilities, even at the expense of over-fitting. The AUC of the ROC curve in the test set decreases from 0.853 to 0.840 by inserting fixed effects in the model, if the model is estimated with ML. This is a natural consequence of over-fitting caused by the large number of players who have not a real decisive role in determining score probabilities of their shots. BR method permits to correct the estimates of all parameters. Specifically, they are shrunk towards zero [2] and, therefore, less variable estimators are produced. This fact reduces the over-fitting and increases the accuracy, with AUC ROC equal to 0.845.

To confirm the suppositions previously expressed, the significance of fixed effects α has been tested. Wilks and Rao tests are not generally available for bias reduced estimators. Nevertheless, since the BR estimator is asymptotically normally distributed [5], Wald test can be used. In particular, it confirms that fixed effects are jointly not significant. But this fact is no longer true if parameters related to player with less than 35 shots (60% of shooters) in the season are fixed to zero. In other terms, it seems clear that a sufficiently large number of shots would allow to decrease latent heterogeneity and to interpret in a proper way the idiosyncratic error. However, stratification by players was also considered, above all, to allow the composition of a shooting ability ranking by ordering fixed effects. In this respect, Figure 1 shows the most interesting results for players with large number of shots during the season. Fixed effects summarise more information than the commonly used conversion rates, that are identified in the plot by the positions of the points with respect to the red line. Unfortunately, the lack of data about defenders' position affects to a greater extent some players, as strikers for instance, who risk more shots under strong defensive pressure.

As for the general covariates, not surprisingly, *distance* and *angle* of shots are shown to be fundamental variables to determine the scoring probability. They are

both characterized by a decreasing polynomial behaviour on $g(\pi)$. Furthermore, also the *type of shot* and the *type of assist* influence the outcome. Specifically, it seems generally unlikely to score shots subsequent to passes. A more detailed interpretation of this effect is given in the following.

If the link function is defined as the complementary log-log function, then a single unit increase of the k -th covariate produces a variation of a factor $e^{\hat{\beta}_k}$ on the quantity $-\log(1 - \pi)$. A simple expansion gives $\pi \approx -\log(1 - \pi)$ for small values of π . Since around 75% of the distribution of predicted π is concentrated on the interval $(0, 0.15)$, the proposed approximation seems reasonable. Hence, let us consider $\hat{\beta}_H = -1.088$ and $\hat{\beta}_C = -1.636$ as the estimated parameter related to covariates indicating an header shot and a cross assist respectively. It is immediate to verify that the probability to score a header after a cross is $e^{-1.088-1.636} \approx 0.066$ times the probability to score a regular shot from the same position. Moreover, the variable representing the interaction between *distance* and *cross indicator* has an impact of 11% on the effect of the *distance*, confirming the idea introduced by Caley [1], according to which a different model should be specified for headers.

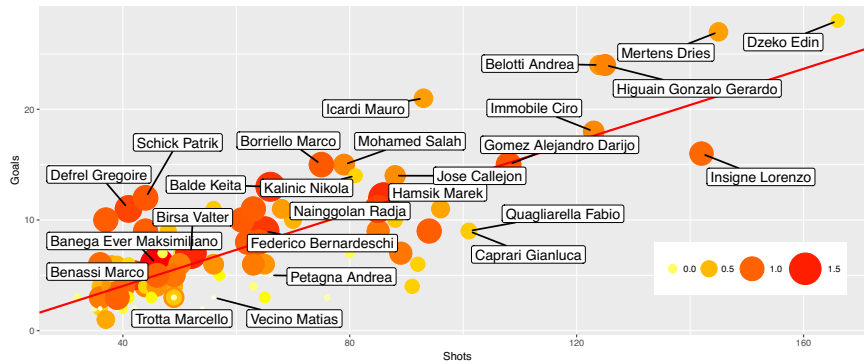


Fig. 1 Size and colour of the points are scaled according to the estimates of fixed effects. On the red line there are players with conversion rates equal to the average conversion rate.

4 Discussion

A complementary log-log binary regression model has been applied to assess scoring probabilities of shots. The use of BR method has provided estimates with good properties, also in the case of a fixed effects specification. Testing and model comparison have been performed using Wald statistics, but Score statistics, while asymptotically equivalent, could be an alternative worth exploring, given that they are often more accurate in models for discrete data (see for instance [14]).

Lack of data about defenders' position has been observed as a strong limitation of this work, since it could strongly affect any type of conclusion. For instance, some goal chances could be wrongly evaluated, even if, as shown by predictive performance in Sect. 3, the model presents an overall satisfying accuracy. As it has been

already mentioned, also fixed effects estimates display drawbacks and sources of systematic bias. Despite of this fact, such estimates of individual abilities represent a valuable information to assess players and, therefore, to take strategic decisions in a football club. The dissemination of advanced computer vision systems will allow to obtain more informative data and hopefully to improve the strength of results.

Two possible extensions represent focus of future work. Firstly, only the attacking actions concluded by shots have been considered in this paper; instead, with more information, the outcome of every pitch situation could be modelled as goal/no-goal. Secondly, the modelling of the entire generating process of a goal, by a decomposition in sub-models (as shot frequency, dangerousness and XG model), would allow to obtain the probability mass function of the number of goals in advance with respect to the match. Moreover, by considering fixed effects model for each part, it would be possible to assess every player in different attacking abilities.

Finally, even with the limitations underlined above, mostly due to lack of information in the available data, the proposed procedure highlights the potential of statistical tools in providing useful and easily accessible solutions to sports insiders.

Acknowledgements The authors are grateful to Lazar Petrov for providing the access to InStat data and to Jacopo Diquigiovanni for advice and comments.

References

1. Caley, M.: Premier League Projections and New Expected Goals. In: Cartilage Free Captain (2015). <http://cartilagefreecaptain.sbnation.com/2015/10/19/9295905/premier-league-projections-and-new-expectedgoals>. Cited 15 Jun 2018
2. Firth, D.: Bias reduction of maximum likelihood estimates. *Biometrika*. **80**, 27–38 (1993)
3. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Springer series in statistics, New York. (2001)
4. In: InStat (2018). <http://www.instatsport.com/>. Cited 15 Jun 2018
5. Kosmidis, I.: Bias in parametric estimation: reduction and useful side-effects. *Wiley Interdiscip. Rev Comput Stat*. **6**, 185–196 (2014)
6. Kosmidis, I.: brglm2: Bias Reduction in Generalized Linear Models, R package version 0.1.7. In: CRAN-R Project (2018). <https://CRAN.R-project.org/package=brglm2/>. Cited 15 Sep 2018
7. Kosmidis, I., Kenne Pagui, E., Sartori, N.: Mean and median bias reduction in generalized linear models. *Stat Comput*, <http://doi.org/10.1007/s11222-019-09860-6>. (2019)
8. Lunardon, N.: On bias reduction and incidental parameters. *Biometrika*. **105**, 233–238 (2018)
9. McHale, I., Szczepański, Ł.: A mixed effects model for identifying goal scoring ability of footballers. *J R Stat Soc Ser A*. **177**, 397–417 (2014)
10. Neuhaus, J.M., McCulloch, C.E.: Separating between-and withincluster covariate effects by using conditional and partitioning methods. *J R Stat Soc Ser B*. **68**, 859–872 (2006)
11. Neyman, J., Scott, E.L.: Consistent estimates based on partially consistent observations. *Econometrica*. **16**, 132 (1948)
12. Pollard, R., Reep, C.: Measuring the effectiveness of playing strategies at soccer. *J R Stat Soc Ser D*. **46**, 541–550 (1997)
13. Sartori, N.: Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*. **90**, 533–549 (2003)
14. Siino, M., Fasola, S., Muggeo, V.M.: Inferential tools in penalized logistic regression for small and sparse data: A comparative study. *Stat. Methods Med. Res.* **27**, 1365–1375 (2018)
15. Wright, C., Atkins, S., Polman, R., Jones, B., Sargeson, L.: Factors associated with goals and goal scoring opportunities in professional soccer. *Int J Perf Anal Spor*. **11**, 438–449 (2011)

Looking for Efficient Methods to Collect and Geolocalise Tweets

Alla ricerca di metodi efficienti per raccogliere e geolocalizzare tweet

Stephan Schlosser, Daniele Toninelli and Silvia Fabris

Abstract In the era of Big Data, Internet became one of the main data sources: data can be collected for relatively low costs and can be used for a wide range of purposes. The target of this paper is to compare three alternative methods of Twitter posts collection. We evaluate three main criteria: the accuracy, the relative reliability and the relative efficiency of each method, using first NUTS level of Great Britain data. One of the three methods stands as being the best (reusable) compromise.

Abstract Nell'era dei Big Data, Internet si è rivelata una delle fonti di dati principali: i dati vengono raccolti ad un costo relativamente basso e possono essere utilizzati per i più svariati scopi. Obiettivo principale di questa ricerca è confrontare tre metodi alternativi per raccogliere post inviati via Twitter. I criteri di valutazione a cui ricorriamo sono tre: l'accuratezza, l'affidabilità e l'efficienza relativa di ciascun metodo, usando dati riferiti al primo livello di NUTS della Gran Bretagna. I risultati indicano che uno dei tre metodi emerge come miglior compromesso.

Key words: social network, big data, geotagging, tweet analysis, data quality.

1 Introduction

Our society is producing, constantly, a huge amount of data, sometimes without even being aware about it. In this era of “Big data”, research focused on collection methods, or aimed at measuring and addressing data quality issues is playing a

¹ Stephan Schlosser (University of Göttingen), stephan.schlosser@sowi.uni-goettingen.de; Daniele Toninelli (University of Bergamo), daniele.toninelli@unibg.it; Silvia Fabris (University of Bergamo), silvia.fabris@unibg.it.

fundamental role. Internet became one of the main (and still largely unexplored) data source: it represents the new frontier, providing data that can be retrieved from social networks or by means of web-scraping techniques. Data available through the web can be collected for relatively low costs, they are immediately available, they can be almost automatically stored in digital formats and they can be used for a wide range of purposes. Nevertheless, using Internet can also be challenging: these types of data are frequently unstructured, most of the times they are not collected for the specific purposes of the research and they are huge in dimension, which needs proper transformations and intensive computational effort.

This paper focuses on testing three alternative methods (introduced in sect. 3) that aim at collecting and geolocalise messages (tweets) sent through the Twitter social network, in order to allow researchers to study socio-economic phenomena and to estimate their trends at a very detailed level, from the geographical point of view. Moreover, this will allow to update, for relatively reduced costs and times, indicators or results obtained from large-scale surveys or by agencies producing official statistics. Once selected, the proposed tool will be, as much as possible, detailed (based on small territorial units, called circles), accurate (fully covering all studied sub-areas) and wide (able to collect as many data/tweets as possible).

We will choose the best method, among the alternatives, according to three criteria (see sect. 4): accuracy in terms of geolocalisation, relative reliability (in covering the different sub-areas and in reflecting the territorial distribution of the reference population) and relative efficiency (producing complete, but also reasonably sized datasets, from the computational point of view). Our attention will be focused on Great Britain (GB), particularly at the first NUTS level (sub-areas)¹.

2 Literature review

Collection and geolocalisation of tweets became crucial for many applications, such as real-time event detection, sentiment analysis, natural disaster analysis and transportation planning (Paule, 2018). Nevertheless, only 1-2% of tweets provide geographic coordinates and user self-reported location is not always reliable. To improve the geolocalisation accuracy, Paule et al. (2018) suggest to consider different features of each individual geotagged tweet (e.g. number of hashtags/mentions/urls in the text, day of the week, etc.). Using a machine learning strategy, for each new tweet the N-most similar geotagged tweets are identified and then a majority voting algorithm is used to select the most common area (of size 1 km).

Han et al. (2014) geolocalise tweets at the city level by means of a multi-class text classification algorithm based on user-declared metadata and “location indicative words”. These words either explicitly (e.g., place names) or implicitly (e.g., dialectal words, slang) contain geographical information.

In comparison to what done previously, our research aims at collecting data already-geolocalised (since the geolocalisation is a direct product of the collection

¹ For a definition of NUTS, see: <https://ec.europa.eu/eurostat/web/nuts/background>.

methods). This implies that we collect tweets via the (official) Twitter API¹, which allows to collect tweets from a specific geographical area (circle). To ensure that the tweets provided by the Twitter API originate from the specific region, tweets containing geographic information (GPS coordinates) are verified for accordance with this previously defined geographical area. If this is the case without exception, it can at least be assumed that the other tweets recorded also originate from this geographical region. This would allow tweets to be treated and analysed separately according to different sub-areas of countries (NUTS levels).

3 Data collection

The data we used were collected from January 15th to February 15th 2019 using three PCs, in order to apply, in parallel, three different collection methods aimed at collecting and geolocalising tweets sent all over the GB. All methods rely on the “theory of circles”, that aims at covering as much accurately as possible the regions of interest by means of a set of circles (differently or equally sized).

The first method (**M1**, from here on) is an enhanced version of the original manual method of collection introduced in Toninelli, Cameletti & Schlosser (2018), set in order to improve the accuracy in covering the GB NUTS sub-areas. This method takes into account three groups of areas: a) big cities, i.e. cities with more than 350,000 habitants (each of them is represented by a unique circle); b) London (divided into 73 small circles); c) other areas (covered by one or more circles that should represent, geographically, each NUTS).

The second method (**M2**) is semi-automatic² and sets circles covering the different sub-areas using a radius that is inversely proportional to the density of the population observed in that sub-area. In particular, a radius of 5 km is set for highly populated regions, 10 km circles are set for moderately populated regions and radius bigger than 10 km are set for less populated areas (the radius is specifically set by sub-area); 3 km circles are used for the city of London. This second method might attenuate the problem of not complete coverage of M1. Moreover, we expect M2 to be more accurate than M1, since bigger circles can be acceptable in low-density regions and because M2 allows a semi-automatic set of circles over the country.

The third method (**M3**) relies on the use of small and equally sized circles (3km of radius for London, 10 km of radius for the other sub-areas). This third method eases the reproduction of NUTS borders and is expected to reduce also the overlaps between NUTS sub-areas. Moreover, it is easier to be set up, since it allows to create an automatic algorithm that could be later replicated in other countries.

The three methods, from a geographical perspective, are compared in Figure 1.

¹ <https://help.twitter.com/en/rules-and-policies/twitter-api>.

² The geographical areas are defined according to guidelines that could be reused for the future.

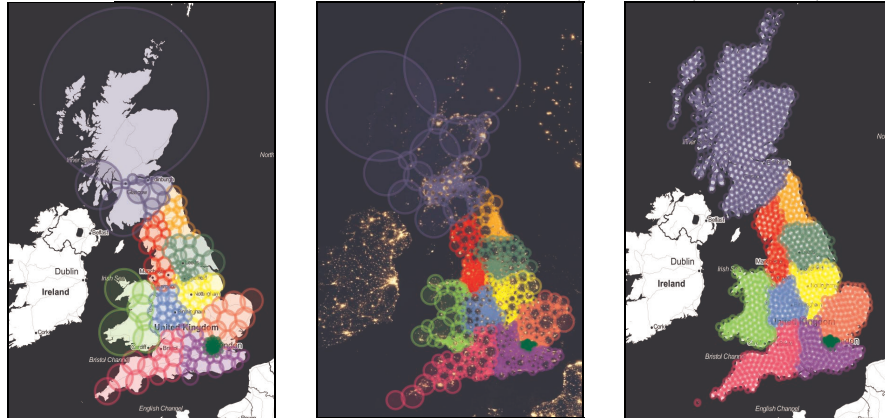


Figure 1: The “theory of circles” applied according to the methods **M1** (left), **M2** (center), **M3** (right)

In total, we collected about 120 millions of tweets, corresponding to about 250 GB of memory storage and to about 750 hours of data collection (details are shown, by method, in Table 1). Seen the frequent overlap between circles set for the collection, tweets collected twice were, first, detected (basing on the text string/id) and then randomly reassigned to one of the two overlapping circles.

4 Methodology

In order to compare the three alternative methods, we take into account three main general factors: accuracy, coverage and efficiency. **a) Accuracy:** the accuracy of collected data (in terms of geolocalisation) is evaluated studying the percentage of tweets that has a geocode (*a.1*) and the share of non-overlapping tweets (i.e. tweets that have not to be reassigned to one of two overlapping areas) (*a.2*). **b) Coverage:** we evaluate the relative reliability of each method studying how much the collected tweets are able to reflect the population distribution density estimated for different NUTS. This is done according to two criteria: the number of NUTS for which each method shows the best performance (lowest difference between density of tweets and population density) (*b.1*); the average percentage difference, by NUTS, between the tweets density and the population density (*b.2*). **c) Computational efficiency:** we compare the three methods in terms of storage size (*c.1*), collecting times (*c.2*) and computational time needed to process tweets for some basic tasks (e.g. remove BOT, URL, punctuation and replace emoji, slang and money symbols) (*c.3*).

5 Results

Table 1 shows the main results for the comparison factors introduced previously.

Table 1: Comparison of three methods by comparison factor

<i>Comparison factor</i>	<i>M1</i>	<i>M2</i>	<i>M3</i>
About data collection			
No. of circles used	246	847	1,257
Tweets amount (in millions)	38.49	43.10	37.92
a) Accuracy			
a.1 Tweets with a geocode	1.03	1.01	1.05
a.2 Share of non-overlapping tweets	94.43	93.60	90.92
b) Coverage			
b.1 NUTS with lowest difference	4	3	4
b.2 Density % diff. (tweets vs popul.)	32.88	32.37	33.74
c) Efficiency			
c.1 Data stored (million unique tw.)	36.35	40.33	34.51
c.2 Collect. time (mins per 10k unique tw.)	3.97	3.93	4.26
c.3 Processing time (mins)	14.32	15.05	15.82

According to the capability of reducing the effort in setting the data collection, the best option seems to be, at first sight, M1. For this method, the number of circles to be set (246) is noticeably lower than the number of circles needed for M2 and M3. Nevertheless, the circles setting with M1 is almost manual, thus a very good compromise seems to be M2, where the circles setting is semi-automatic. Moreover, M2 with a big, but still reasonable, number of circles (847; 32.6% less than the ones asked by M3) is also able to maximise, generally, the number of collected tweets (12.2% more than M1 and 13.7% more than M2); the latter seems to suggest a better and more complete coverage of the different sub-areas.

For what concern **accuracy (a)**, when studying the geotagged tweets, with GPS coordinates available, we notice that all methods show an almost perfect attributions of tweets to the specific sub-areas of reference (except very few cases that may be attributed to the random reassignment of overlapping tweets). In terms of share of overlapping tweets (i.e. tweets that are collected for two different circles), the best results (i.e. the lower percentage) are obtained using M1 (5.2%). These findings can be explained by means of the presence of very big circles for the areas excluding big cities (that lower a lot the overlap, in comparison to M3) and/or by the noticeable reduction of the amount of circles used, in comparison to both M2 and M3. However, also according to this point of view, an excellent compromise seems to be M2, that lowers the risk of the “border” effect (for big circles the tweet collection is not able to fully cover the reference sub-area). Moreover (detail not shown in Table 1), M2 is able to collect the highest number of unique tweets in 8 out of 11 NUTS.

Evaluating **coverage (b)**, results are mixed and are strictly dependent upon the considered NUTS. According to the number of NUTS for which a method shows the best performance (lower difference between tweets and population densities), the two best modes are M1 and M3 (with four NUTS each), whereas M2 is the best method in 3 cases out of 11. Nevertheless, according to the average percentage

difference, the best method seems to be M1 (5.6%), showing an average percentage difference lower than both M2 (6.4%) and M3 (9.1%).

In terms of **efficiency (c)**, M2 results as the best method of collection, bringing to the biggest amount of stored data, after the reassignment (40.33 million of unique tweets) and also to the fastest collection time (3.9 mins requested to deal with 10k unique tweets). Moreover, considering the processing time of 10k tweets, M2 seems to do not differ much from the other methods (15.05 secs in comparison to 15.82 necessary for data collected with M3 and to 14.32 necessary with M1).

6 Conclusions, limits and further research

M2 seems to be superior to the other methods both in terms of the requirements needed to set up the collection procedure and of the total number of unique tweets collected, as well as for its efficiency for most of the processing activities. For all the other comparison factors, M2 represents a good compromise that compensate the non-automatic feature of M1 and the excess of time needed to collect and reassign the tweets observed with M3. Nevertheless, for some indicators (e.g. for the percentage difference between the tweets and the population density) results are mixed, suggesting a further and deeper study of the performances of the three methods, considering more indicators and a higher level of detail (at least set at the different NUTS 1 level). In addition to this, further research is needed in order to obtain a better understanding of the relationship between the amount of collected tweets and the population characteristics in the different NUTS regions.

We developed our study involving just GB, which being an island can affect the collection methods: in other countries, circles can overlap borders, making less reasonable the random reassignment of overlapping tweets. Moreover, in order to use a wider range of data, collection is still on and we planned to check the efficiency of the three methods also performing more advanced studies such as using the sentiment analysis. Finding out the most efficient and reliable method, we aim at enhance the data collection method itself, and we project to expand our collection, involving more countries, in order to use this strategy to measure and estimate socio-economic trends, with a multi-national perspective.

References

1. Han, B., Cook, P., Baldwin T.: Text-Based Twitter User Geolocation Prediction. In: Journal of Artificial Intelligence Research, 49, 451-500 (2014)
2. Gonzalez Paule, J.D., Sun, Y., Moshfeghi, Y.: On fine-grained geolocalisation of tweets and real-time traffic incident detection. J. Information Processing & Management. (2018, in press) doi: 10.1016/j.ipm.2018.03.011
3. Toninelli, D., Cameletti, M., & Schlosser, S.: New Frontiers in Measuring the Well-being in the Big Data Era. In: 2018 JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association. 2094-2109 (2018) Available via <http://hdl.handle.net/10446/132220>.

Principal ranking profiles

Principal ranking profiles

Mariangela Sciandra, Antonella Plaia

Abstract Descriptive statistics and methods for exploratory analysis of data widely introduced in literature are useful in presenting an overall picture of data. Not only do they provide a summary of the data, but they are also often suggestive of the appropriate direction to analyse the data. It is a common practice to begin an analysis by examining graphical displays of a data to get a better sense of the data to be analyzed. Ranking (and rating) data, which result from n raters ranking (rating) m items, are difficult to visualize due to their discrete algebraic structure, and the computational difficulties associated with them when n is large. This problem becomes worse when raters provide tied rankings or not all items are ranked. This work is devoted to the definition of a new approach for the visualisation of ranking (rating) data for large n that allows to identify the principal rater preference profiles.

Abstract *Statistiche descrittive e metodi per l'analisi esplorativa di dati ampiamente introdotti in letteratura sono utili per presentare un quadro generale dei dati. Non solo forniscono un riepilogo dei dati, ma spesso suggeriscono anche la direzione appropriata per analizzare i dati stessi. È prassi comune iniziare un'analisi partendo dalla rappresentazione grafica degli stessi al fine di ottenere una migliore comprensione dei dati da analizzare. I dati espressi in forma di classifica, che provengono dalle valutazioni di n giudici che classificano m oggetti, sono difficili da visualizzare a causa della loro struttura algebrica discreta, e le difficoltà computazionali che ne derivano quando n è grande. Questo problema diventa ancora più complesso quando i valutatori forniscono una classifica che prevede parimerito o quando non tutti gli oggetti sono classificati. Questo lavoro è dedicato alla definizione di un nuovo approccio per la visualizzazione dei dati di ranking per n di grandi dimensioni che consente di identificare i principali profili di preferenze del valutatore.*

Key words: Ranking data, Preference profiles, Preference plot, PCA.

Mariangela Sciandra, Antonella Plaia
University of Palermo, Viale delle Scienze, Building 13, 90128, Palermo - Italy.
e-mail: mariangela.sciandra@unipa.it / antonella.plaia@unipa.it

1 Introduction

To display a set of rankings, it is not advisable to use traditional graphical methods such as histograms and bar graphs because the elements of R , the set of all possible permutations of the m objects, do not have a natural linear ordering. Visualization techniques for ranking data have drawn the attention of many researchers. Some of the methods introduced in literature are basically some generalization of classical graphical methods for quantitative data while some are specifically defined in the context of ranking data. When ranking data are analyzed common aims of the graphical representation should be to identify the typical ranking of the m objects (the consensus ranking) but also to evaluate the agreement among the judges. Moreover, another interesting topic concerns the ability of the plot in identifying possible outliers among the judges and/or the objects. Note that when the size of the ranking data is large (e.g., $m > 5$ or $n > 100$), it is practically impossible to reveal the above-mentioned pattern and characteristics by merely looking at the raw data or by using some simple descriptive statistics such as the means and standard deviations of the ranks.

Geometrically, rankings of m objects can be represented as points in \mathbb{R}^{m-1} . The set of all $m!$ rankings can then form a convex hull of $m!$ points in \mathbb{R}^{m-1} known as a permutation polytope. The idea of using a permutation polytope to visualize ranking data was first proposed by Schulman (1979). Other graphical representations for ranking data use multidimensional scaling (Marden, 1995) to plot a set of judge points in a low-dimensional Euclidean space or some unfolding techniques (Coombs, 1950) able to visualize a set of points in a low Euclidean space with both judges and objects being represented by the points in the same space. In this section, we introduce a new graphical approach to display the important features of a large collection of preference profiles by applying a generalization of the methodology introduced by Jones and Rice (1992) to the case of ranking data. So, after a brief remind on ranking data, in the third section the Principal profiles plot is introduced based on the Principal Component Analysis of judge profiles. A real dataset example and some ideas for future work conclude this paper.

2 Preference data

Ranking data arise when a group of judges is asked to rank a fixed set of objects (*items*) according to their preferences. When ranking k items, labeled $1, \dots, m$, a ranking π is a mapping function from the set of items $\{1, \dots, m\}$ to the set of ranks $\{1, \dots, m\}$, endowed with the natural ordering of integers, where $\pi(i)$ is the rank given by the judge to item i . When all m items are ranked in m distinct ranks, we observe a complete ranking or *linear ordering* (Cook, 1986). Yet, it is also possible that a judge fails to distinguish between two or more objects and assigns them equally, thus resulting in a tied ranking or *weak ordering*. Besides complete and tied rankings, *partial* and *incomplete rankings* exist: the first occurs when only a

specific subset of $q < m$ objects are ranked by judges, while incomplete ranking occurs when judges are free to rank different subsets of m objects. Obviously, different types of ordering will generate different sample space of ranking data. With m objects there are $k!$ possible complete rankings; this number gets even larger when ties are allowed.

3 How to display ranking data: the Principal Profile plot

Graphs are the most important tool for investigating ranking data because they allow a comparative look among individual profiles that neither tables or summary statistics cannot ever bring out. Modern high-speed computing allowed recent developments in displaying ranking data. One particularly useful plot could be obtained as a hybrid plot combining a scatterplot of the rankings with a line drawing connecting rankings from the same judge. Figure 1 contains an example of this hybrid plot: on the horizontal axis there are the 12 different items while on the vertical axis scores attributed to each item by subjects are represented. In this toy example a different colour for each subject is used in order to show how confusing is its interpretation when the number of subjects is high.

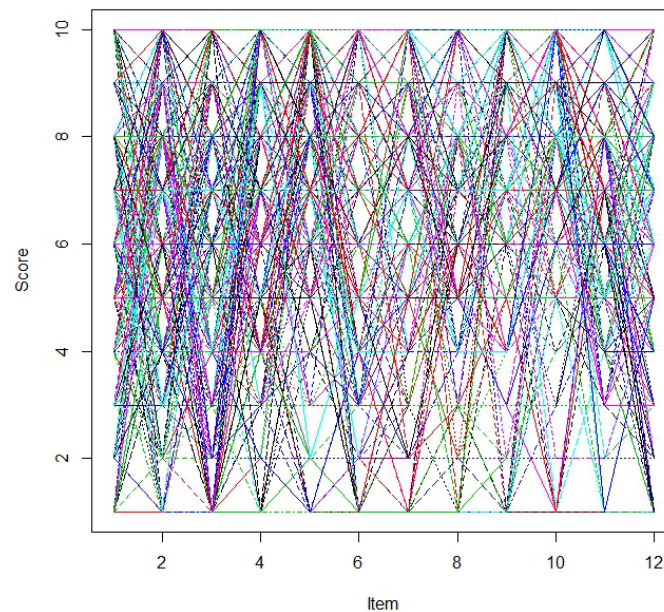


Fig. 1 Confusion plot of individuals ratings on 12 items of a questionnaire.

The aim is to use the Principal components in order to identify representative curves from a collection of curves. Their solution is to treat each of the N curves as a vector of p function values evaluated on an equispaced grid. Thus a collection of N observations on p variables is obtained. Principal component analysis is applied to the p variate data set in order to identify representative curves. Once principal components are obtained, representative profiles are extracted by using the coordinates of each unit in the new coordinate system (principal component scores). In particular, once all the units are ranked on the basis of their scores on the first principal component, the underlying criterion for profile selection will consist into select profiles corresponding to the five Tukey's magic numbers of the distribution of the profile scores on the first mode of variation. In other words, we assume that an hypothetical profiles distribution exists and we extract units summarising this unknown distribution. In this way, the selected profiles will not be a combination of the original profiles but a subset of the original observed set. In order to be sure about the goodness of selected profile in summarising the underlying distribution we fix the $100\alpha\%$ quantile as $p([N] + 1)$, where p_r is the r -th order statistics of the unit distribution according to the principal component scores and $[x]$ is the integer part of x . The points selected are then transformed into coordinates of the original curves and graphed. The resulting Principal Ranking Profile plot (PRP-plot) obtained by applying the proposed technique to the plot in Fig1 is shown in Fig2:

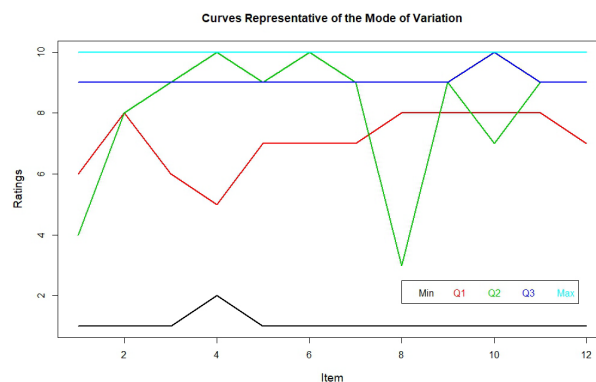


Fig. 2 PRP-plot on a real dataset

4 Conclusion

Through a PCA on a group of several profile a Principal Profile plot is proposed. The results show our graphical method allows to identify the typical response profile when rating and ranking data are used. A goal for the future will be the the use of the

Principal ranking profiles

proposed graphical tool as a way to identify the consensus profile and to fix some constraints to solve the problem of crossing profiles.

References

1. Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, 57(3), 145-158.
2. Jones, M. C., and Rice, J. A. (1992). Displaying the Important Features of Large Collections of Similar Curves. *The American Statistician*, 46, 140-145.
3. Marden, John I.(1995). *Analyzing and modeling rank data*. London : Chapman and Hall.
4. Robert S. Schulman (1979). Ordinal data: An alternative distribution. In: *Psychometrika*, 44(1), pp. 3–20.

A statistical model for voting probabilities

Un modello statistico per le probabilità di voto

Rosaria Simone, Stefania Capecchi

Abstract In the last decades, the proportion of voters who express their vote according to their specific needs –and not following some ideology– is increasing, not only in Italy. This circumstance represents a prominent factor for political (in)stability and implies that common surveys where people are asked about their “intentions to vote” should be accurately examined to extract any useful information about political consensus. In fact, these intentions change with time and circumstances fairly often and thus it may be suitable to investigate preferences towards the whole spectrum of political options to acquire more definite evaluations. Thus, distributions of voting probabilities require parametric modelling to infer on the polarization and indifference components of each candidate (party, leader, coalition, etc.). Some empirical evidence seems to support the proposed approach.

Abstract Negli ultimi decenni è cresciuta una componente elettorale che –non essendo riferita ad una scelta ideologica ben definita– si orienta di volta in volta verso la persona/leader che, al momento, meglio rappresenta le proprie esigenze. Ciò genera una mutovelezza costante per cui l'analisi dei sondaggi rivolti alla cosiddetta “intenzione di voto” deve essere perseguita nella sua globalità. A tal fine, ci si può riferire a modelli capaci di spiegare le componenti caratterizzanti le probabilità di voto per ogni candidato (partito o leader), e cioè l'indifferenza e la polarizzazione, per esempio. Le evidenze empiriche presentate sembrano confermare l'utilità dell'approccio proposto.

Key words: Political surveys, Voting probabilities, Preference data models, Indifference, Polarization

Rosaria Simone and Stefania Capecchi

Department of Political Sciences, University of Naples Federico II, Naples, Italy. e-mail: rosaria.simone@unina.it, stefania.capecchi@unina.it

Name of Second Author

Name, Address of Institute e-mail: name@email.address

1 Introduction

Surveys aiming to investigate the consensus of public opinion towards governments and political parties or leaders are frequent and prominent, also to predict electorate behaviour and to accommodate programs and objectives of stakeholders. In this area, several techniques have been introduced by pollsters to collect “voting probabilities” as suitable predictors of the final intention of the electorate. This approach is often limited to a single option and it is effective when voters have a well defined idea and the political preference is a consequence of a precise ideology.

By contrast, a larger and larger proportion of people moves their vote among parties according to temporary psycho-social sentiments, job related problems and economic issues. If this description is correct, it is more important to investigate the evaluation of the whole scenario of parties (in terms of leaders, coalitions, etc.) since this appraisal reflects personal judgements with higher fidelity. Such an approach would reveal itself as particularly appealing to measure the polarization of the choices, their heterogeneity as well as the importance of indecision and indifference in the evaluation process.

In this paper, a modelling framework will be exploited to infer on the main components of the voting probabilities towards political parties. Specifically, the proposal is aimed to disentangle polarization, indecision and indifference of the choices when respondent is asked to assess the probability to vote a given party.

After an introduction to the proposed model, the paper its performances on empirical survey data and suggests further research directions.

2 The methodological framework

Let us assume that, for a given m , the response R to the question-item:

“On the scale 1 to m , how likely is it that you would ever vote for party A?”

is a discrete random variable whose observations –in a sample survey of size n – are $\mathbf{r} = (r_1, \dots, r_n)'$.

In this kind of interview, two patterns are to be expected: bimodal distributions with modes at the extreme categories (when polarization is evident) or a unique modal value at one of the extremes (when supporters/opponents are dominant). In fact, interviewees may be adverse (opponents), indecisive (that is, indifferent, uncertain with the choice) or favourable to party A (supporters), with different intensity.

Then, a mixture distribution of two components may be introduced to explain intrinsic patterns of voting probabilities:

- A proportion δ of interviewees shows a “neutral” (indifferent) behaviour and prefers to take refuge at the midpoint $s = (m + 1)/2$ of the support of R , where

for convenience we assume m is an odd integer. The distribution of this group is degenerate at $R = s$, that is $D_r^{(s)} = I(r = s)$, $I(\cdot)$ being the indicator function.

- The dual proportion $1 - \delta$ of respondents expresses a graduated preference or aversion towards A , with modal values at $R = 1$ (the opponents) and/or $R = m$ (the supporters), respectively. In this respect, the discretized Beta random variable $db_r(\alpha, \beta)$, defined over the support $\{1, 2, \dots, m\}$, and characterized by parameters $\alpha > 0$ and $\beta > 0$ may well fit such a pattern (more details in [5, 6, 2] and [4]) if $\min(\alpha, \beta) < 1$. The difference $\alpha - \beta$ indicates the shape of the distribution (U-, J- or bell- shaped).

Formally, the discretized Beta random variable $X \sim db_r(\alpha, \beta)$ is derived from discretization of a continuous Beta distribution over m sub-intervals of the support $(0, 1)$ of equal length, and it is defined as:

$$Pr(X = r | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_{\frac{r-1}{m}}^{\frac{r}{m}} x^{\alpha-1} (1-x)^{\beta-1} dx, \quad r = 1, 2, \dots, m,$$

$\Gamma(z)$ being the Gamma function for any $z > 0$. The distribution of X is symmetric when $\alpha = \beta$ and it collapses to a discrete Uniform distribution if $\alpha = \beta = 1$, that is when interviewees randomly select the response to avoid any burden with respect to a political choice (or there is maximum heterogeneity of voting intentions). Properties of discretized Beta derive from those of the continuous counterpart: in particular, positive, null and negative skewness occur for $\alpha < \beta$, $\alpha = \beta$ and $\alpha > \beta$, respectively. J-shaped (or reverse J-shaped) distributions are for $\alpha \geq 1$; $\beta < 1$ (or $\alpha < 1$; $\beta \geq 1$); U-shaped distributions, with modal values at the extremes, occurs when $\alpha < 1$ and $\beta < 1$. Thus, the probability distribution of voting probability R is:

$$Pr(R = r | \boldsymbol{\theta}) = \delta D_r^{(s)} + (1 - \delta) db_r(\alpha, \beta), \quad r = 1, \dots, m, \quad (1)$$

where $\boldsymbol{\theta} = (\alpha, \beta, \delta)'$. The Discretized Beta with Inflation (DBI) model (1) requires $m > 4$, with parameters establishing patterns of respondents' voting intentions. When analysis is performed with reference to subgroups, then it allows to emphasize differential behaviours. Thus, in order to address campaign interventions and to identify drivers of voting behaviours, one should consider covariates' effect by linking individual characteristics' values $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w}_i)$ to parameters:

$$\text{logit}(\delta_i) = \mathbf{x}_i \boldsymbol{\omega}; \quad \log(\alpha_i) = \mathbf{y}_i \boldsymbol{\gamma}; \quad \log(\beta_i) = \mathbf{w}_i \boldsymbol{\eta}. \quad (2)$$

When $\alpha = \beta = 1$, model (1) becomes a CUSH model [1] to fit data characterized by indifferent (who select midpoint) and indecisive responses.

Given the sample information $\mathbf{C}_n = (r_i, \mathbf{x}_i, \mathbf{y}_i, \mathbf{w}_i, i = 1, \dots, n)$, maximum likelihood (ML) estimates $\boldsymbol{\theta} = (\alpha, \beta, \delta)$ are obtained by maximizing the log-likelihood:

$$\ell(\boldsymbol{\theta}, \mathbf{C}_n) = \log L(\boldsymbol{\theta}, \mathbf{C}_n) = \sum_{i=1}^n \sum_{r=1}^m I(R_i = r) \log Pr(R_i = r | \boldsymbol{\theta}, \mathbf{C}_n).$$

Based on the asymptotic efficiency of ML procedures, statistical tests and goodness of fit measures are introduced to check for the validity of the estimated model.

A feature of the proposed discrete model for voting probabilities is that it is possible to derive measures of polarization and indifference for each candidate:

- Straightforwardly, the estimated $\hat{\delta}$ is a measure of *indifference*, neutrality.
- The relative distance of supporters with respect to opponents (the extent to which supporters outclass opponents) for a given political candidate, can be assessed via the normalized measure in $[-1, 1]$:

$$\rho(\hat{\alpha}, \hat{\beta}) = (1 - \hat{\delta}) \left(db_m(\hat{\alpha}, \hat{\beta}) - db_1(\hat{\alpha}, \hat{\beta}) \right).$$

In case of perfect polarization (that is, for symmetric distributions), $\rho(\alpha, \beta) \approx 0$.

- The strength of polarization of voting probabilities can be assessed in terms of:

$$s_p(\hat{\alpha}, \hat{\beta}) = 2(1 - \hat{\delta}) \max \{0, q(\hat{\alpha}, \hat{\beta})\},$$

where

$$q(\hat{\alpha}, \hat{\beta}) = \min\{db_1(\hat{\alpha}, \hat{\beta}), db_m(\hat{\alpha}, \hat{\beta})\} - \max\{db_2(\hat{\alpha}, \hat{\beta}), \dots, db_{m-1}(\hat{\alpha}, \hat{\beta})\}.$$

Such normalized measure in $[0, 1]$ is positive only in case of *U-shaped* distributions with modal values at the extremes (not necessarily symmetric), otherwise it is zero. In that case, there basically exists a unique pole (of opponents or supporters depending on the sign of $\rho(\alpha, \beta)$). If positive, a higher value indicates a stronger polarization for the given candidate.

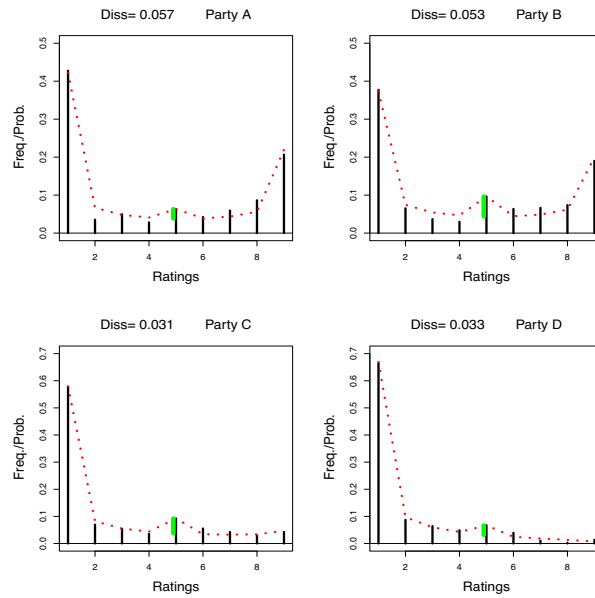
These measures can be considered in a temporal perspective for a given candidate to derive a feedback from the political electorate and foresee voting intentions; in the following, we set $\hat{\rho} = \rho(\hat{\alpha}, \hat{\beta})$ and $\hat{s}_p = s_p(\hat{\alpha}, \hat{\beta})$ for short.

3 Some empirical evidence

We discuss the performance of model (1) by exploiting some recent opinion surveys on Italian political parties based on stratified samples of Italian adults. A group of interviewees have been asked in January 2019 to assess the probability to vote for a party according to a 10-point rating scale; then, the scale is reduced to 9 scores by considering that both 9 and 10 evaluations coincide *de facto* with the maximum $m = 9$. Two time occasions t_0 and t_1 will be considered, corresponding to two groups of subjects. For simplicity, voting probabilities for the main parties (coded as A, B, C, D) of last Italian political elections have been analysed. Table 1 lists the main features of the estimated models (1), whereas Figure 1 highlights the comparison between observed and fitted distributions (dissimilarity is also reported).

Table 1 Voting probability models for the main parties (standard errors in parentheses)

Parties	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\rho}$	\hat{s}_p	$Diss$
A (t_0)	0.025 (0.010)	0.190 (0.016)	0.294 (0.024)	-0.204	0.305	0.057
A (t_1)	0.025 (0.009)	0.177 (0.014)	0.280 (0.022)	-0.215	0.316	0.063
B (t_0)	0.052 (0.012)	0.239 (0.019)	0.351 (0.028)	-0.173	0.252	0.053
B (t_1)	0.053 (0.011)	0.248 (0.019)	0.387 (0.028)	-0.198	0.205	0.082
C (t_0)	0.055 (0.012)	0.183 (0.017)	0.723 (0.064)	-0.533	0	0.031
C (t_1)	0.019 (0.009)	0.196 (0.016)	0.765 (0.060)	-0.546	0	0.037
D (t_0)	0.035 (0.011)	0.208 (0.022)	1.421 (0.153)	-0.662	0	0.033
D (t_1)	0.043 (0.010)	0.205 (0.020)	1.505 (0.149)	-0.671	0	0.035

**Fig. 1** Observed and estimated distributions of voting probabilities at time occasion t_0 . The bold segment at midpoint emphasizes the estimated indifference proportion

Both estimation and plots confirm the adequacy of the proposed model. In particular, one can remark that:

- indifference, as measured by $\hat{\delta}$, is quite limited for parties A and D, whereas for B and C it is larger than 5%; however, at time t_1 , indifference for party C drops to less than 2%;

- After controlling for indifference, for parties A and B , there is an asymmetrical polarization, which is stronger for party A (as measured by \hat{s}_p), with a prevalence of opponents, which is also stronger for party A (as indicated by \hat{p}).
- After controlling for indifference, for parties C and D there is no polarization and there exists a unique pole, composed of opponents.
- Polarization for party A is stronger at t_1 than t_0 : given that the distance between supporters and opponents increases (considering the absolute value of \hat{p}), one can conclude that at time t_1 voting probabilities were more resolute. For party B , instead, if on one hand the distance between supporters and opponents has also increased (again, considering the absolute value of \hat{p}), polarization as measured by \hat{s}_p has decreased: this means that voting probabilities at t_1 showed more indecisive patterns than at t_0 , with a polarization that is weaker, and more right-asymmetrical than at time t_0 .

4 Discussion

The proposed model is tailored to parameterize voting probabilities when expressed on a rating scale. Given its flexibility to encompass different shapes, its application lends itself to the introduction of simple measures for voting polarization and indifference. Having available voting probabilities on different time occasions, such indexes could serve to monitor voting intentions. A further development will concern a comparative analysis with a mixture of IHG (Inverse HyperGeometric) distributions, to identify clusters with divergent opinions in rating surveys [3].

Acknowledgements: This work has been partially supported by CUBREMOT project, University of Naples Federico II. Data courtesy of Noto Sondaggi S.r.l.

References

1. Capecchi S. and Piccolo D. (2017). Dealing with heterogeneity in ordinal responses. *Quality & Quantity*, **51**, 2375–2393.
2. Fasola S., Sciandra M. (2015) New Flexible Probability Distributions for Ranking Data. In: Morlini I., Minerva T, Vichi M (eds.), *Advances in Statistical Models for Data Analysis*. Springer, Springer-Verlag, pp.117–124
3. Simone R. and Iannario M. (2018) Analysing sport data with clusters of opposite preferences, *Statistical Modelling*, 18(5–6), 505–524.
4. Simone R. and Tutz G. (2018) Modelling uncertainty and response styles in ordinal data. *Statistica Neerlandica*, 72: 224–245
5. Ursino M. (2014) Ordinal Data: a new model with applications, *PhD Thesis*, <http://porto.polito.it/2535701/>, Politecnico di Torino, Italy.
6. Ursino M. and Gasparini M. (2018) A new parsimonious model for ordinal longitudinal data with application to subjective evaluation of a gastrointestinal disease, *Statistical Methods in Medical Research*, 27(5), 1376–1393.

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

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

Gloria Solano-Hermosilla¹, Fabio Micale², Vincenzo Nardelli³, Julius Adewopo⁴, Celso Gorrín González⁵

Abstract

Agriculture is the main source of income in African countries, but there is an untapped potential to support informed decisions or optimize returns on investment to market actors due to imperfect distribution of market signals, among other factors. In Africa the spread of smartphones has given rise to Citizen Science projects to collect and share data. Yet, data quality has received minimal attention. We present a smartphone-based approach to collect and disseminate near real-time commodity food prices in Nigeria from spontaneous crowd volunteers. We applied a stepwise quality approach to overcome relevant statistical issues and produce reliable price estimates. The results highlight the potential of this method and open new paths in the digital era for robust spatial-temporal analyses of prices and market behaviours.

Abstract

L'agricoltura è la principale fonte di reddito nei paesi africani, ma esiste un potenziale di produzione non sfruttato, dovuto, tra l'altro, alla distribuzione imperfetta dei segnali tra gli attori del mercato che rende più difficile per gli agricoltori prendere decisioni informate per massimizzare il reddito e aumentare la capacità di investimento. In Africa la diffusione degli smartphone ha dato vita a progetti Citizen Science al fine di raccogliere e condividere i dati. Tuttavia, occorre controllare la qualità dei dati raccolti. In questo paper presentiamo un'App per smartphone sviluppata al fine di raccogliere e diffondere in tempo reale prezzi di beni alimentari rilevati in forma volontaria (crowdsourcing) in Nigeria. Nel paper presentiamo un approccio qualitativo graduale al fine di superare statistiche relative a campioni e non campioni. I risultati evidenziano le potenzialità del metodo e aprono nuovi percorsi nell'era digitale per l'analisi spaziale dei prezzi e dei comportamenti di mercato.

Key words: Validation, post-sampling, post-stratification, crowdsourcing, Citizen Science, smartphone app, food prices

1 Introduction

Agriculture is a main source of income in African countries; however, there is an untapped production potential [1] due, among others, to the imperfect distribution of market signals across market actors along the food supply chain. Timely and reliable

¹ Gloria Solano-Hermosilla, Joint Research Centre, European Commission; email: gloria.solano-hermosilla@ec.europa.eu

² Fabio Micale, Joint Research Centre, European Commission; email: fabio.micale@ec.europa.eu

³ Vincenzo Nardelli, Università Cattolica del Sacro Cuore, Milano (IT); email: Vincenzo.nardelli01@icatt.it

⁴ Julius Adewopo, International Institute for Tropical Agriculture (IITA); email: J.Adewopo@cgiar.org

⁵ Celso Gorrín González, Joint Research Centre, European Commission; email: Celso.gorri-gonzalez@ec.europa.eu

market information can empower farmers and help them to make informed decisions to maximize income, increasing so their welfare and investment capacity [2]. It also provides an important source to inform national and regional decisions and policies regarding food security and markets. The current lack of timeliness and spatial representation of various market segments may limit opportunities for supply chain interventions to strengthen actors and bolster food security. Thus, increasing the availability of market information in Africa has risen in recent years to the top of the agendas of both national statistical offices and development partners.

In Africa the spread of mobile- and smartphones has given rise to multiple Citizen Science¹ projects (e.g. crowdsourcing² initiatives) and applications to collect and share data, searching to empower participants of the agricultural supply chain (e.g. farmers, processors, traders and consumers) by connecting them as both providers and users of collective and valuable information [3].

We find examples of crowdsourcing projects in which participants are engaged to submit real-time food price data via mobile phone from several public initiatives (e.g. from World Bank, mVAM-WFP, JRC, AfDB, FAO-AMIS) in combination or not with private companies that have explored in Nigeria and/or other African countries different types of citizen engagement (e.g. spontaneous vs. recruited participants), rewards (e.g. monetary vs. phone airtime) and technologies (classic vs. smartphone) [4], [5], [6], [7], [8], [9], [10]. These initiatives show the potential to improve timeliness and coverage of areas of difficult access or too costly for traditional data collections. Yet, data need to be of good quality; they must be accurate, representative, sustained in time and accessible at a speed that meets the needs of governments, private institutions, and commodity supply chain actors.

Crowdsourcing needs to deal with the spontaneity and intermittency of volunteers, often anonymous and of unknown and variable expertise and reliability [11]. The anonymity may also encourage cheating to maximize income. Crowdsourced data are derived from non-probabilistic, convenient samples of self-selected individuals willing to participate [12]. Reliability, accuracy and representativeness become major concerns. As advantages, crowdsourcing provides a quick way to obtain and analyse food market prices and is a cost-efficient means to increase the sample with more frequent data on a larger geographic scale. It may also facilitate a more efficient two-way communication, fostering market orientation in agriculture. A main challenge, however, is the inference of the true price [13]. Research on this topic proposes several solutions but often not comparable or not easily applicable to different datasets or a diversity of tasks types [13], [11].

This paper suggests how Citizen Science and smartphones can be used to collect and disseminate accurate food price data in real time. We first present a smartphone app and a crowdsourcing method to collect prices in Nigeria along the agricultural supply chain. We fine tune and apply a quality approach to overcome the statistical problems of crowdsourced data, which do not obey a formal sampling design and may suffer other non-sampling problems, to produce timely and reliable estimates [12]. Finally, we present an online dashboard to make reliable prices daily accessible. This research opens new paths in the digital era for spatial analysis of food prices and market behaviours that could be expanded to other data domains or other type of digital data collection (e.g. scan data, web scraping).

2 The “Food Price Crowdsourcing in Africa” data collection

The “Food Price Crowdsourcing in Africa” (FPCA) is joint collaboration of the European Commission's Joint Research Centre (EC-JRC), the International Institute

¹ Citizen Science refers to “projects in which volunteers partner with scientists to answer real-world questions”, in particular, but not only, to expand opportunities for data collection and to provide access back to scientific information for community members [20].

² Crowdsourcing is a technique by which citizens conduct a task, e.g. collect data, using ICTs on a purely voluntary basis, with no control a-priori on the sampling design, the data collection strategy and its quality. Crowdsourcing can be considered as a Citizen Science initiative.

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 of Tropical Agriculture (IITA) and Wageningen University and Research (WUR), to understand food price changes while enhancing agricultural & market information systems through mobile phone technology and people's collaboration. Since September 2018 the FPCA project is implementing a crowdsourcing approach where after an advertisement campaign, anonymous volunteers spontaneously engage to submit data at any day/time on prices of selected food commodities (i.e. rice (local and imported), maize, beans and soybean), from different markets along the food chain (i.e. farm gate, wholesale and retail markets) with focus in Kano and Katsina States in the Northern region of Nigeria. Crowd volunteers (VCs) must be in possession of a smartphone with GPS functionality and able to follow online instructions. No training is provided; it is a quasi "contactless" initiative. The system is built based on open data kit (ODK) and deployed on a compatible cloud-based server, ONA, which assimilates data submissions in real-time. The reward system is designed to be "non-committal" and yet "promising" for the crowd members. Based on daily, weekly, and monthly thresholds, VCs whose submissions are valid are rewarded on a "first-submit, first-rewarded" basis. In addition, behavioural or psychological factors (known as "nudges") are mobilised in the design of the approach to help sustain data submissions (see JRC, 2019, forthcoming, for a full description of the methodology).

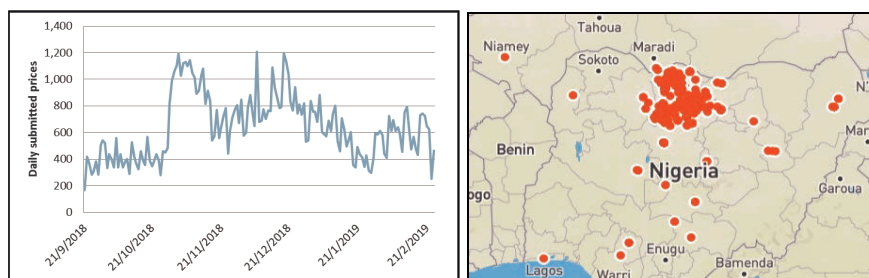


Figure 1: Daily number of price data submissions (left) and map of submissions (right).
Source: FPCA project

After around 5 months, the on-boarded ~800 VCs have submitted ~12,500 individual data records, corresponding to more than 100,000 price observations. Following the initial 6-weeks pilot phase with an average 46 daily submissions (~288 prices) with VCs ~200, we moved into the roll-out phase with an average of ~100 daily submissions. The peak submissions were observed around weekends (Friday – Sunday), with maximum submissions (n=168) recorded in the 7th week. A behavioural "social norm" nudge was implemented to promote sustained submissions and reversed the declining trend started in week 14. This consisted of a non-binding weekly communication after an initial 21 week period without any feedback and is showing positive results with submissions peaking one day after sending the weekly communication.

In terms of crowd profile, it is currently composed of registered 643 males and 144 females. Most of the volunteers are between the ages of 20-35 year, and nearly all have attained tertiary or secondary school education. Breakdown of the crowd by profession shows that a majority of volunteers are either students (~40%) or directly engaged in the agricultural sector (36%). Generally, 4 out of every 5 volunteers have more than 4 years' experience with using smartphones.

The registration form is very useful to identify most promising opportunity for engagement and retention of the on boarded crowd. The majority of volunteers knew about the initiative through the distributed handbill (~71%), the rest of volunteers from friends or neighbours (~18%) or agricultural extension agents (~9%), and to less extend from internet or radio (2%). Over 70% of the volunteers indicated that they are interested in regular data feedback and more than 80% indicated to prefer bank transfer for reward.

3 Quality methodology and results

Crowdsourcing can be applied to solve different types of tasks requiring decision making, selecting options or providing a value. Our approach concerns the last type, which is defined as a numerical task [13] where the VC is asked to provide the price of certain food commodities within a certain geographic area. How can we infer the ‘true’ price values of the area?

Based on the refinement of previous work of Arbia et al. [14], [15], [12] in this section we introduce and apply to our crowdsourced sample a stepwise quality procedure consisting of a series of algorithms to produce reliable, accurate and real-time accessible food price area estimates that allow for statistical inference. We tackle so 3 important dimensions of data quality frameworks, at the base in most National Statistical Institutes [16], [17], and along which we develop our approach and present the results: accuracy and reliability, timeliness and accessibility.

3.1 Accuracy and reliability

It refers to the degree to which data match the phenomena they are designed to measure. The difference to the ‘truth’ can be attributed to many error components that can be divided into ‘sampling’ and ‘non-sampling’ errors.

Sampling problems are connatural to crowdsourced data and derive mainly from the lack of a proper sample design. Non-sampling errors can be related to possible measurement errors due to wrong interpretations of the crowd, time invariance of the prices (e.g. due to the repetition of the same value several times to avoid checking), locational errors (due to mistakes in the recording of the coordinates), possible fraudulent activities (e.g., those submitting duplicate reports using multiple accounts). Some of the possible non-sampling errors we tackled in the data collection methodology by, for example, setting rules to avoid duplicated crowd profiles and minimizing the number of fields that require manual entry (e.g. closed list of most common packaging units). Once this is settled, we propose a stepwise quality approach consisting of several algorithms¹ (Figure 2) to tackle non-sampling (step 0 to 2) and sampling errors (step 3) to produce reliable area price estimates, which lead to consecutive datasets with increased level of quality check accompanied by specific quality control metrics:

Step 0- Raw Crowdsourced Data are the individual price observations provided by each VC without any editing or another form of processing but anonymizing of fields with personal information.

Step 1- Primary Crowdsourced Data are the Raw Crowdsourced Data (individual observations) after some preliminary processing mainly related to editing, automatic conversion to standard measurement units, geolocation to administrative divisions, i.e. State, Local Government Area (LGA) and Ward-urban/rural.

Step 2 –Processed Crowdsourced Data are the Primary Crowdsourced Data (individual observations) gone through the validation process based on outlier detection techniques. In crowdsourcing for numerical tasks the truth inference is usually based on the aggregation of multiple observations from multiple contributors for the same task, e.g. collect the price of 1 kg rice in a certain market (“Crowd wisdom” [18]). We propose a procedure with customizable parameters to build first spatial clusters (DBSCAN), and then identify within-cluster price outliers. As a result outliers and isolated points (those that could not find a cluster) are removed and the mean of the remained values can be used to estimate the true price. Although results are available for all commodities on a daily basis, we report here the results of a week for retail prices of local and Indian rice which stay at the two extremes. We obtained 735 (in 26 LGAs and 52 wards) and 89 (in 21 LGAs and 38 wards) observations for local and Indian rice respectively in Kano and Katsina States. We

¹ All algorithms developed in each step are programmed in R software [21] and available by the authors

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 found a 4% and a 39% of outliers and a 2% and a 20% of isolated points for local and Indian rice respectively.

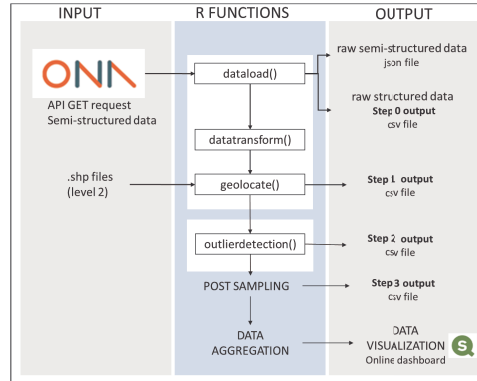


Figure 2: Stepwise description of the quality methodology, developed R functions and output.

Step 3 - Aggregated Crowdsourced Data are the Processed Crowdsourced data after going through the desired temporal aggregation (e.g. daily, weekly) and post-sampled (PS) at the desired level of spatial aggregation (e.g. State). By crowdsourcing spontaneous price contributions from volunteers the lack of a proper sample design leading to problems of extrapolating results to population, thus of representativeness. Here we run a post-sampling, a special form of post-stratification by which crowdsourced data are reweighted prior their use in an inferential context. We produce a formal random stratified sample design proportional to population size for the number of actual crowdsourced observations. The map of observed data is then compared with the map of the formal sample design resulting in calculated weights, the PS ratios. The actual observations are then reweighted in such a way to resemble the formal sampling scheme allowing for statistical inference. We can assess the reliability of the data acquisition process. In fact, if the crowdsourcing coincides perfectly with the desired formal sampling we achieve the maximum of reliability.

In contrast, the larger will be the deviation from 0 (both positive and negative) the lower the reliability. In this sense, a good *intrinsic* measure of the crowdsourcing reliability is given by:

$$CSR = 1 - 2 \left[\frac{\frac{\sum_{i=1}^L (m_i - n_i)^2}{N}}{1 + \frac{\sum_{i=1}^L (m_i - n_i)^2}{N}} \right]$$

Where m_i is the number of observations crowdsourced in location i and n_i are the number of observations of the sampling plan in i . The index Crowdsourcing Reliability (CSR) ranges between 0 and 1. Indeed, when $m_i - n_i = 0, \forall i$, $CSR = 1$ and we have a perfect correspondence between the crowdsourced data and the desired observations thus achieving the highest possible reliability of the survey. Conversely, when $\sum_{i=1}^L (m_i - n_i)^2$ tends to infinity, we have to the lowest possible reliability and CSR index tends to zero. The proposed method may be applied to different geographical levels. In addition we can measure *extrinsically* the validity of the crowdsourced dataset by aggregating the data accordingly to a reference dataset and calculating Root Mean Square Errors (RMSE). The results for local rice show a CSR of 0.7 and low values of RMSE (i.e. in urban areas) when we compare the post-sampled price with data provided by the National Bureau of Statistics (NBS) of Nigeria [19].

3.2 Timeliness and punctuality

It refers to the length of the time gap between the actual collection of the food price data by the participant crowd and the availability of the validated and aggregated data. Our set of algorithms proved to be technically fit to run from one to several times a day and data can be aggregated for each commodity at any needed time frequency and administrative partition.

3.3 Accessibility and clarity

Data need to be timely produced but also made accessible and easy interpretable by users to be effective and usable in decision making processes. For this purpose we have designed a user-friendly online dashboard to make accurate data available in real-time and actionable for data users (Figure 3). Technically the data resulting from the quality algorithms can be automatically uploaded in the dashboard with the desired frequency.

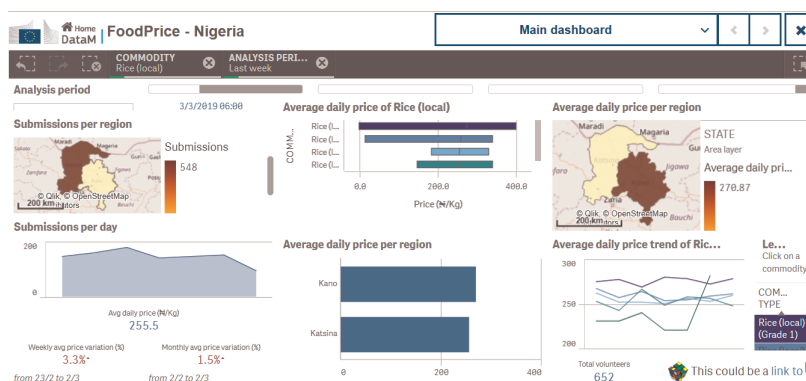


Figure 3: DataM Dashboard – FPCA daily crowdsourced food prices in Nigeria.

4 Conclusions

The paper refers to the FPCA project collecting food prices in Nigeria via a smart phone app and a crowdsourcing approach based on the engagement of spontaneous and anonymous volunteers to provide data without any predefined level of quality or sample design. We propose and fine tune the application of a quality methodology [12] consisting on a series of validation algorithms to produce quality food price data.

The application of the methodology shows that by integrating ICT tools and scientific knowledge we are able to make data available in real-time, accessible, accurate, and actionable for better decision making from farmers up to consumers and support national and regional policy making regarding food security and markets. From a practical perspective the proposed quality methodology can be useful for institutions or organizations that aim at complementing the price data collection systems with crowdsourcing approaches. This research opens new paths in the digital era for spatio-temporal analysis of food prices and food access and to understand market behaviors that could be expanded to other regions, data domains or other type of digital data collection (e.g. scan data, web scraping).

References

1. AGRA, “Africa Agriculture Status Report: The Business of Smallholder Agriculture

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

- in Sub-Saharan Africa (Issue 5)," Nairobi, Kenya, 2017.
2. S. Fan, J. Brzeska, M. Keyzer, and A. Halsema, *From subsistence to profit: Transforming smallholder farms*, vol. 26. Intl Food Policy Res Inst, 2013.
3. FAO, "Agricultural Services and Digital Inclusion in Africa," 2018. [Online]. Available: <http://www.fao.org/family-farming/detail/en/c/1105795/>.
4. N. Hamadeh, M. Rissanen, and M. Yamanaka, "Crowd-sourced price data collection through mobile phones," *Collab. Res. Methodol. Off. Stat. World Bank*. http://ec.europa.eu/eurostat/cros/sites/crosportal/files/NTTS2013fullPaper_82-v2.pdf, 2013.
5. WFP, "2016 - Mobile Vulnerability Analysis & Mapping (mVAM)," 2016. [Online]. Available: <https://www.wfp.org/content/2016-mobile-vulnerability-analysis-mapping-mvam>.
6. A. Donmez *et al.*, "Using web and mobile phone technologies to collect food market prices in Africa, Approaching real-time data and use of crowdsourcing, 2013 – 2016. JRC Working Papers JRC104311, Joint Research Centre (Seville site).," 2017.
7. Standard Chartered, "'CROWDSOURCING' Indicator for Africa Consumer price tracker and business sentiment indicator part of package to address data gaps across region," *Standard Chartered*, no. June, 2014.
8. J. E. Blumenstock and N. Keleher, "The Price is Right?: Statistical evaluation of a crowd-sourced market information system in Liberia," in *Proceedings of the 2015 Annual Symposium on Computing for Development*, 2015, pp. 117–125.
9. Y. Seid and F. Fonteneau, "Leveraging crowdsourcing techniques and technologies to generate better agricultural information: three case studies, in Indonesia, Nigeria and the Philippines. FAO, Statistics Division," 2017.
10. H. Zeug, G. Zeug, C. Bielski, G. Solano-Hermosilla, and R. M'barek, "Innovative Food Price Collection in Developing Countries. Focus on Crowdsourcing in Africa. JRC Working Papers JRC103294, Joint Research Centre (Seville site).," Joint Research Centre (Seville site), 2017.
11. T. J. Bird *et al.*, "Statistical solutions for error and bias in global citizen science datasets," *Biol. Conserv.*, vol. 173, pp. 144–154, 2014.
12. G. Arbia, G. Solano-Hermosilla, F. Micale, and V. Nardelli, "Post-sampling crowdsourced data to allow reliable statistical inference : the case of food price indices in Nigeria di crowdsourced per un ' inferenza statistica più affidabile : il caso degli indici di prezzo in Nigeria," in *49th Scientific meeting of the Italian Statistical Society*, 2018.
13. Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, "Truth Inference in Crowdsourcing : Is the Problem Solved?," in *Proceedings of the VLDB Endowment*, 10(5), 2017, pp. 541–552.
14. G. Arbia, "The use of GIS in spatial surveys," *Int. Stat. Rev.*, vol. 61, no. 2, pp. 339–359, 1993.
15. G. Arbia, *A primer for spatial econometrics*. Palgrave Macmillan, 2014.
16. UN, "Guidelines for the template for a generic national quality assurance framework (NQAF)," 2012.
17. ESS, "Quality Assurance Framework of the European Statistical System Table of Contents," 2015.
18. J. Surowiecki, *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York; London and Toronto, 2004.
19. NBS, "National Bureau of Statistics, Nigeria," 2019. [Online]. Available: <https://www.nigerianstat.gov.ng/>.
20. J. P. Cohn, "Citizen Science : Can Volunteers Do Real Research?," *Bioscience*, vol. 58, no. 3, 2008.
21. R Core Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing." Vienna, Austria, 2013.

Dealing with uncertainty in automated test assembly problems

La gestione dell'incertezza nei problemi di assemblaggio automatizzato dei test

Giada Spaccapanico Proietti, Mariagiulia Matteucci and Stefania Mignani

Abstract The recent development of computer technologies enabled test institutes to improve the test assembly process by automated test assembly (ATA). A general framework for ATA consists in adopting mixed-integer programming models. These models are intended to be solved by common commercial solvers which, notwithstanding their success in handling most of the known problems, are not always able to find solutions for highly constrained and large-sized ATA problems. Moreover, all parameters are assumed to be fixed and known, an hypothesis that is not true for estimates of item response theory (IRT) parameters. In this work, we propose a chance-constrained model for dealing with uncertainty in ATA without increasing the complexity of the model.

Abstract Il recente sviluppo delle tecnologie informatiche ha consentito agli istituti di valutazione di migliorare il processo di assemblaggio dei test tramite l'automated test assembly (ATA). Una struttura generale per ATA consiste nell'adottare modelli di programmazione intera-mista. Questi modelli sono pensati per essere risolti da solver commerciali che, nonostante il loro successo nella gestione della maggior parte dei problemi noti, non sono sempre in grado di risolvere problemi di ATA molto vincolati o di grandi dimensioni. Inoltre, tutti i parametri sono considerati fissi e noti, un'ipotesi che non vale per le stime dei parametri di item response theory (IRT). In questo lavoro proponiamo un modello chance-constrained per affrontare l'incertezza nei modelli di ATA senza aumentarne la complessità.

Key words: Psychometrics, Automated test assembly, Mixed-integer programming, Non-linear programming, Chance constrained programming, Julia

Giada Spaccapanico Proietti, Mariagiulia Matteucci, Stefania Mignani
Department of Statistical Sciences, University of Bologna,
via delle Belle Arti n.41, Bologna.
giada.spaccapanico2@unibo.it, m.matteucci@unibo.it,
stefania.mignani@unibo.it

1 Introduction

In educational measurement, a test is a collection of items developed to measure students' abilities. In order to make different measurements, comparable tests should be standardized i.e. the test procedures must be fixed in such a way that differences among testing conditions do not influence the scores [8]. Test assembly consists in selecting items from an item pool to build parallel test forms that meet some desiderata (specifications). Thus it plays a fundamental role in test development. Several methods are used for test assembly nowadays either based on classical test theory (CTT) or item response theory (IRT).

Larger testing programs, have better access to resources like sophisticated item banking systems, opening the possibility to improve their test assembly process by means of automated test assembly (ATA). These procedures are very important to make measurements comparable while reducing operational costs. In practice the ATA models are not always easy to solve because they involve a very large number of decision variables and constraints. Also with the standard form suggested in [7], it's not possible to define any part of the model in a non-linear manner. Another limit of classical ATA models, until now applied, is that they consider each variable as fixed or known. This is not valid for IRT item parameter estimates, which describe the psychometric properties of the items, and subsequently for the item information function (IIF), which is a key element in ATA models.

These restrictions motivated us to find an alternative way to specify and solve ATA models. We suggest a *chance-constrained* approach, from here CC [see 1] which allows to maximize the $(1 - \alpha)$ -quantile of the empirical distribution function of the test information function (TIF) obtained by bootstrapping the calibration process. In this way we ensure that, independently on the situation in which the calibration has been made, we have an high probability to have a certain error (possibly low) in the ability estimation. For solving the ATA models, CC or not, we applied a stochastic meta-heuristic called *simulated annealing* (SA) proposed by [2]. This technique can handle large-scale models and non-linear functions and has the main feature that can always find a feasible solution (or the most feasible) because it avoids to being trapped in a local optimum thanks to a random variable that accepts or not an inferior solution.

All the algorithms used in this work has been coded in Julia¹ because of its high-performance in numerical analysis and for the availability of many valuable packages for optimization.

2 Automated test assembly models

Van der Linden book [7] provide a general framework for 0-1 Linear (0-1 LP) and mixed-integer (MIP) programming models for automated test assembly. These methods allow to assemble optimal tests with a set of specifications. Once the prob-

¹ <https://julialang.org/>

lem has been specified by using the MIP standards for constraints and objective function it is solved by a software that, hopefully, gives as a result the structure of the test forms that satisfies the specifications. This software may be commercial such as CPLEX or Gurobi or open-source like COIN-OR ² optimization suites. If a set of desiderata is well specified and has a linear representation, those models can always be written in the following standard form:

$$\begin{aligned} &\text{optimize} && c^T x && \text{(objective function)} \\ &\text{subject to} && Ax - b \leq 0 && \text{(constraints),} \end{aligned} \quad (1)$$

where x is a vector of I binary variables if only one test form is assembled. Constraints may define the structure of the test forms such as length and composition in terms of item category or their statistical properties such as the test maximum expected score. The objective function usually involves some statistical feature of the test forms such as the maximization of the TIF as the sum of the single IIF of the chosen items. Only one objective can be optimized at a time; if we have more than one function to optimize some tricks can be applied to transform the objectives into constraints introducing new continuous decision variables and transforming the 0-1 linear model in a mixed-integer one. On the other hand, there is no upper limit for the number of constraints, provided our solver is able to handle them.

Model (1) can be generalized to the case in which we assemble T test forms just by considering $I \cdot T$ decision variables. This modification will exponentially increase the size and complexity of the model, especially if there are constraints on overlaps between test forms. The latter are the most tough specifications to handle in ATA because their linear representation introduce $\binom{T}{2}$ new binary decision variables and $\left[\binom{T}{2} \cdot I \cdot 2 \right] + \binom{T}{2}$ new constraints.

3 Chance-constrained test assembly

When the optimization model introduced in the previous section is applied, all parameters are assumed to be fixed and known. Unfortunately this assumption doesn't hold for all the variables. IRT item parameters, for example, must be estimated bringing a certain level of uncertainty in the model. So far, uncertainty has been mainly disregarded in modeling test assembly problems, even though the scientific and professional community is aware of it. Optimization problems involving uncertainty occur in almost all areas of science and engineering such as management of electrical systems, finance and medicine. This stimulated the interest in building rigorous ways of formulating, analyzing, and solving such problems giving birth to the theory of stochastic programming [see 9]. Moreover these factors motivated some researchers to address the role of stochastic programming in test assembly models. Two attempts in this direction have been made in [8] and [9] in which robust optimization is applied.

² <https://www.coin-or.org/>

In the context of stochastic programming another way to deal with uncertainty in optimization is the CC approach. This method, even today, has widespread application, due to its ease of use and low demand on computational power compared to other stochastic programming formulations.

Bootstrapping the calibration of item parameters

The strength of IRT models, theoretically, is to have invariant item parameters across samples of examinees from the same population. Practically, invariance is hard to be guaranteed under the calibration process. In [6] is showed that using estimates of item parameters instead of their true values could lead to biases in the following inference about the students' ability. Although existing methods for handling uncertainty in item parameters provide a variety of tools, most of them belong to Bayesian applications, which need to know the prior distribution of item parameters and abilities. An alternative approach that simulates the calibration under different conditions and that doesn't ask to specify any starting distribution is the bootstrap. In particular we performed the calibration in a large number of sufficiently big sub-samples of the test responses with the aim to reconstruct the uncertainty related to each item parameter by exploring its empirical distribution function.

Given an item bank of items and a sample of students, from here called "full sample", after the items have been administered we have a matrix of dichotomous responses. For each item we must estimate a vector $\hat{\xi}_i$ of IRT parameters. The overall calibration (on the full sample of students) and each single repetition is made by following the classic joint-maximum likelihood estimation method (JMLE) [3]. We also add a constraint in each maximization to force the expected score to be not so far from the observed score. For solving the optimization model we used the suite NLOpt³ by the Julia package NLOpt.jl⁴ with the algorithm SLSQP. Once we got the overall estimates $\hat{\xi}_i$ of the item parameters and $\hat{\theta}_n$ of the abilities of the students we proceeded with the bootstrap scheme:

Algorithm 1 Bootstrapped calibration

Choose a large number of repetitions R , the subsample size $N^* < N$ and an ability point θ_0 in which maximize the TIFs.

Discretize the distribution of the ability θ by dividing its continuum in K bins and approximate the probability of sampling the n th student by the relative frequency \hat{p}_n of their $\hat{\theta}_n$ on the full sample.

for $r = 1 : R$ **do**

 Use the vector of \hat{p}_n to sample N^* rows from the responses matrix.

 Calibrate the items in the subsample. Get $\hat{\xi}_r = \{\hat{\xi}_{1r}, \dots, \hat{\xi}_{Ir}\}$.

 Compute the IIF for each item for a chosen set of ability points θ_0 using $\hat{\xi}_{ir}$.

 Store the vectors $\hat{\text{IIF}}(\theta_0)_r = \{\hat{\text{IIF}}(\theta_0)_{1r}, \dots, \hat{\text{IIF}}(\theta_0)_{Ir}\}$

end for

³ <http://ab-initio.mit.edu/nlopt>

⁴ <https://github.com/JuliaOpt/NLOpt.jl>

The stored IIFs are then used to build the empirical distribution function of the TIFs for all the tests to assemble.

The optimization model

Once we have the bootstrapped IIFs we can input them in the following model for the assembly of T test forms:

$$\begin{aligned} & \text{maximize} && w \\ & \text{subject to} && f(x) = \mathbb{P}[\eta^T x_t \geq w] \geq 1 - \alpha \quad \forall t \quad (\text{one-sided linear CC}), \end{aligned} \quad (2)$$

where x_t is a vector of length I , for all $t = 1, \dots, T$, and η is the vector of random variables associated to the IIFs computed at a θ_0 point for all the items. α can be arbitrarily small. These constraints force the TIF of each t to be higher than w with a probability more than $(1 - \alpha)$. Note that for any feasible solution x^* , $f(x^*)$ is an integral of a density function over a polyhedral set so it must be transformed in order to make it computationally tractable. For this aim the sample average approximation (SAA) approach [see 5] is applied.

Let η_1, \dots, η_R be an independent identically distributed sample of R realizations of the random vector η_r (of size I in our case) and P_R be the respective empirical measure hence assigning probability $1/R$ to each point η_r . The SAA $\hat{p}_R(x)$ of a function $p(x)$ is obtained by replacing the true distribution P by the empirical measure P_R . That is $\hat{p}_R(x_t) := P_R(\eta^T x_t \geq w)$. Let $\mathbb{1}_{(w, \infty)} : \mathbb{R} \rightarrow \mathbb{R}$ be the indicator function on (w, ∞) . We can write

$$\hat{p}_R(x) = \mathbb{E}_{P_R}[\mathbb{1}_{(w, \infty)} \eta^T x_t] = \frac{1}{R} \sum_{r=1}^R \mathbb{1}_{(w, \infty)} \eta_r^T x_t \geq 1 - \alpha \quad \forall t$$

where the η_r are the $\hat{\text{IIF}}(\theta_0)_r$ stored in the bootstrapped calibration. The latter will ensure that in the $(1 - \alpha)$ percentage of the calibration repetitions the test information function for all the T test forms will be higher than the maximum value possible w . Of course this method can be generalized effortlessly to the case of shaping the TIFs in more ability points and it is possible to consider two-sided constraints for forcing a percentage of the mass of the considered random variables to be inside a certain range.

4 Simulation study

A simulation study is conducted to investigate the performance of the CC model for ATA problems. We performed the simulation considering the 2-parameters logistic model (2PL). An item bank with $I = 250$ items is simulated, the related IRT discrimination and difficulty parameters $\xi_i = \{a_i, b_i\}$ are sampled respectively from a LogNormal(0,0.25) and a standard normal distribution. We assembled 6 pre-tests

which must be parallel in terms of average approximated difficulty ranging from 1 to 3 (low, medium, high) and with 50 items each. The pre-test design is unbalanced with adjacent forms overlap with 10 items. The responses of $N = 3000$ respondents with abilities θ_n coming from a standard normal are then simulated discarding rows of responses that presented a score too far from the expected one. The overall calibration process started with a run on all the rows of the response matrix. The log-likelihood of the classic JMLE method is maximized using the package NLOpt.jl of Julia framework and the outcoming average RMSEs of the estimates are 0.18, 0.15 and 0.15 respectively for the discrimination, difficulty parameters and abilities of the students. The model has been solved in less than 1 minute with sufficiently good results.

Once the overall abilities estimates are obtained, they are used to select the $R = 500$ subsamples of responses data. The $\hat{I}\hat{F}_{ir}(\theta_0)$ are computed generating a $I \times R$ matrix. Then a model like (2) is solved using the SA algorithm described in [2] under different specifications. The resulting optimal w^* are higher than the ones obtained by the classical model solved with CPLEX 12.8.0⁵ that doesn't consider the uncertainty in IRT parameters, ensuring that with an high probability the error in ability estimation cannot be higher than a certain value.

References

- [1] Charnes, Abraham, and William W. Cooper. Chance-constrained programming. *Management science* 6.1 (1959): 73-79.
- [2] Goffe, William L. SIMANN: a global optimization algorithm using simulated annealing. *Studies in Nonlinear Dynamics & Econometrics* 1.3 (1996).
- [3] Hambleton, Ronald K., Hariharan Swaminathan, and H. Jane Rogers. *Fundamentals of item response theory*. Vol. 2. Sage, (1991).
- [4] Steven G. Johnson, The NLOpt nonlinear-optimization package, <http://ab-initio.mit.edu/nlopt>
- [5] Luedtke, James, and Shabbir Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization* 19.2 (2008): 674-699.
- [6] Tsutakawa, Robert K., and Jane C. Johnson. The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika* 55.2 (1990): 371-390.
- [7] Van der Linden, Wim J. *Linear models for optimal test design*. Springer Science & Business Media, (2006).
- [8] Verschoor, A. J. . *Genetic Algorithms for Automated Test Assembly*. Arnhem: Stichting Cito Instituut voor Toetsontwikkeling, (2007).
- [9] Kall, Peter, Stein W. Wallace, and Peter Kall. *Stochastic programming*. Chichester: Wiley, (1994).

⁵ <https://www.ibm.com/analytics/cplex-optimizer>

Joint Models: a smart way to include functional data in healthcare analytics

Modelli congiunti: un metodo per includere i dati funzionali nelle analisi in ambito sanitario

Marta Spreafico, Francesca Ieva

Abstract The adoption of HealthCare Utilization (HCU) databases allowed to collect heterogeneous data and to improve the quality of healthcare service. In healthcare analytics, associating functional time-varying responses (e.g. biomarkers) with an event-time outcome related to a specific event of interest (e.g. treatment failure or death) represents a challenging task that could be tackled using *Joint Models for longitudinal and time-to-event data*. In particular, the dynamics of consumption of drugs that patients assume during a therapy can be reconstructed through pharmaceutical administrative registries and used to investigate how it influences patients' survival outcomes.

Abstract *L'impiego di databases clinici ed amministrativi ha permesso di ottenere dati eterogenei e di migliorare la qualità del servizio sanitario. Nella ricerca clinica, associare una risposta funzionale che varia nel tempo (ad es. biomarcatori) ad un evento specifico di interesse (e.g. il fallimento del trattamento o la morte) è un problema stimolante che può essere affrontato con modelli congiunti di dati longitudinali e di tempo all'evento. In particolare, la dinamica del consumo di farmaci che i pazienti assumono durante la terapia pu essere ricostruita attraverso i registri amministrativi farmaceutici ed utilizzata per investigare come essa influenzi la sopravvivenza dei pazienti.*

Key words: Joint Models, Time-varying covariates, Healthcare Analytics, Real-World data, Drugs consumption

Marta Spreafico
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: marta.spreafico@polimi.it

Francesca Ieva
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: francesca.ieva@polimi.it

1 Introduction

In recent years healthcare research moved from pursuing only the evidence arising from clinical trials to the Real World (RW) evidence evincible from administrative sources and registries. The use of computers, mobile devices, wearables and other biosensors allowed to gather and store huge amounts of health-related data. These RW data are collected into HealthCare Utilization (HCU) databases [1] with no design and/or clinical questions in mind and they include heterogeneous characteristics: administrative and demographic information, diagnosis, treatments, drugs prescriptions, laboratory tests, hospitalizations or many others.

RW data are particularly suitable for investigating different areas, such as profile of drug uses, post-marketing studies on treatment effectiveness, safety concerns or cost-effectiveness analyses [1]. In healthcare analytics, associating functional responses (e.g. biomarkers) with an event-time outcome related to a specific event of interest (e.g. treatment failure or death) represents a challenging task that could be tackled using *Joint Models (JM) for longitudinal and time-to-event data* [2]. In recent years there has been an increasing use of JM in a wide-range of clinical settings [3] leading to subject-specific predictions and personalized medicine. In particular, in this work we want to reconstruct the dynamics of consumption of drugs that patients assume during a therapy through pharmaceutical administrative registries in order to investigate how it influences patients' survival outcomes.

2 The dataset

In the Lombardia Region dataset patients hospitalized for Heart Failure (HF) from 2000 to 2012 are considered, as described in [4]. For our work, we focus on ACE-Inhibitors therapy and we use a representative sample composed by 2,916 patients, identified by their unique anonymous ID codes and with their first HF hospitalization between 2006-2012. Enrolment occurred from data of discharge of first HF hospitalization (i.e. index date) and the administrative censoring date is December 31st, 2012. Moreover, each record in the dataset is related to an event, which can be a hospitalization or a pharmacological prescription. In the first case, the dates of admission and discharge, together with the length of stay in hospital are given. In the second one, ATC codes, dates of prescription and coverage days are provided.

At index hospitalization, median age is 74 years (IQR 66-80) with a substantial proportion of male patients (57.6%) and the median of total number of comorbidities [5] is 2 (IQR 1-3). During a median follow-up of 51.02 (IQR 33.11-68.21) months, 718 (24.6%) patients die. Finally, since we want to establish if ACE consumption during the first year of follow-up influences patient's survival, we compute a monthly time-dependent variable which indicates the cumulative months covered by ACE assumption up to time t (expressed in months).

3 Joint Models for Longitudinal and Time-to-Event data

In 2010 Rizopoulos proposed Joint Models (JM) for dealing with time-to event and internal time-dependent covariates [2] and wrote the associated R packages JM [6] and JMbayes [7]. In particular, package JMbayes fits a wide range of joint models under a Bayesian approach using Markov Chain Monte Carlo (MCMC) algorithms, allowing the implementation of several types of extensions with respect to JM.

3.1 Submodels specification

Let T_i^* denotes the true event time for the i -th subject, C_i the censoring time, $T_i = \min(T_i^*, C_i)$ the corresponding observed event time and $\delta_i = I(T_i^* \geq C_i)$ the event indicator, with $I(\cdot)$ being the indicator function that takes the value 1 when $T_i^* \geq C_i$, and 0 otherwise. Let $\mathcal{D}_n = \{T_i, \delta_i, \mathbf{y}_i; i = 1, \dots, n\}$ denotes a sample from the target population, where \mathbf{y}_i is the $n_i \times 1$ longitudinal response vector for the i -th subject, with element y_{il} denoting the value of the longitudinal outcome taken at time point t_{il} , $l = 1, \dots, n_i$.

To accommodate different types of longitudinal responses in a unified framework (e.g. continuous normally distributed / categorical / dichotomous / censored longitudinal outcome), Rizopoulos in [7] postulates a generalized linear mixed effects model. In particular, the conditional distribution of \mathbf{y}_i given a vector of random effects \mathbf{b}_i is assumed to be a member of the exponential family, with linear predictor given by

$$g[E\{y_i(t)|\mathbf{b}_i\}] = \eta_i(t) = \mathbf{x}_i^T(t)\boldsymbol{\beta} + \mathbf{z}_i^T(t)\mathbf{b}_i \quad (1)$$

where $g(\cdot)$ denotes a known one-to-one monotonic link function, $y_i(t)$ denotes the value of the longitudinal outcome for the i -th subject at time point t , $\boldsymbol{\beta}$ is the vector of the unknown fixed effects parameters, \mathbf{b}_i is the vector of random effects, $\mathbf{x}_i(t)$ and $\mathbf{z}_i(t)$ denote the time-dependent design vectors for the fixed and random effect, respectively. Moreover, the random effects \mathbf{b}_i are assumed to follow either a multivariate normal distribution with mean zero and variance-covariance matrix \mathbf{D} or a multivariate Student's- t distribution with mean zero, scale covariance matrix \mathbf{D} and df degrees of freedom. For the survival process, Rizopoulos assumes that the risk for an event depends on a function of the subject-specific linear predictor $\eta_i(t)$:

$$\begin{aligned} h_i(t|\mathcal{H}_i(t), \boldsymbol{\omega}_i) &= \lim_{dt \rightarrow 0} \frac{Pr\{t \leq T_i^* < t + dt | T_i^* \geq t, \mathcal{H}_i(t), \boldsymbol{\omega}_i\}}{dt} \\ &= h_0(t) \exp\{\boldsymbol{\gamma}^T \boldsymbol{\omega}_i(t) + f(\mathcal{H}_i(t), \mathbf{b}_i, \boldsymbol{\alpha})\}, \quad t > 0 \end{aligned} \quad (2)$$

where $\mathcal{H}_i(t) = \{\eta_i(s), 0 \leq s < t\}$ denotes the history of the underlying longitudinal process up to time point t , $h_0(\cdot)$ denotes the baseline hazard function, $\boldsymbol{\omega}_i$ is a vector of exogenous, baseline or possibly time-varying, covariates with corresponding

regression coefficient $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ is a parameter vector that quantifies the association between features of the marker process up to time t and the hazard of an event at the same time point. Various options for the form of function $f(\cdot)$ are listed in Table 1. Moreover, the logarithm of the baseline hazard function $h_0(\cdot)$ is modelled using a B-splines approach with vector of spline coefficients $\boldsymbol{\gamma}_{h_0}$.

Table 1 Association structure for the survival process (2) implemented by Rizopoulos in [7].

$f(\mathcal{H}_i(t), \mathbf{b}_i, \boldsymbol{\alpha})$	The risk depends on
$\alpha\eta_i(t)$	the current underlying value at time t
$\alpha\eta'_i(t) = \alpha \frac{d\eta_i(t)}{dt}$	the slope of the current value at time t
$\alpha_1\eta_i(t) + \alpha_2\eta'_i(t)$	both the current value and its slope at time t
$\alpha \int_0^t \eta_i(s) ds$	the area under the longitudinal trajectory up to time t (cumulative effect)
$\alpha \int_0^t \bar{\omega}(t-s)\eta_i(s) ds$	the weighted cumulative effect up to time t
$\boldsymbol{\alpha}^T \mathbf{b}_i$	the random effect only

3.2 Parameters estimation

The estimation of the JM parameters proceeds under a Bayesian approach, using MCMC algorithms. In particular, the expression for the posterior distribution of the model parameters is derived under the assumptions that given the random effects, both the longitudinal and event time process are assumed independent, and the longitudinal responses of each subject are assumed independent:

$$p(\mathbf{y}_i, T_i, \delta_i | \mathbf{b}_i, \boldsymbol{\theta}) = p(\mathbf{y}_i | \mathbf{b}_i, \boldsymbol{\theta}) p(T_i, \delta_i | \mathbf{b}_i, \boldsymbol{\theta}), \quad (3)$$

$$p(\mathbf{y}_i | \mathbf{b}_i, \boldsymbol{\theta}) = \prod_l p(y_{il} | \mathbf{b}_i, \boldsymbol{\theta}) \quad (4)$$

where $\boldsymbol{\theta}$ denotes the full parameter vector, and $p(\cdot)$ denotes an appropriate probability density function. Under these assumptions the posterior distribution is analogous to:

$$p(\boldsymbol{\theta}, \mathbf{b}) \propto \prod_{i=1}^n \prod_{l=1}^{n_i} p(y_{il} | \mathbf{b}_i, \boldsymbol{\theta}) p(T_i, \delta_i | \mathbf{b}_i, \boldsymbol{\theta}) p(\mathbf{b}_i | \boldsymbol{\theta}) p(\boldsymbol{\theta}), \quad (5)$$

where

$$p(y_{il} | \mathbf{b}_i, \boldsymbol{\theta}) = \exp \left\{ \frac{y_{il} \psi_{il}(\mathbf{b}_i) - c\{\psi_{il}(\mathbf{b}_i)\}}{a(\boldsymbol{\varphi}) - d(y_{il}, \boldsymbol{\varphi})} \right\} \quad (6)$$

with $\psi_{il}(\mathbf{b}_i)$ and $\boldsymbol{\varphi}$ denote the natural and dispersion parameters in the exponential family, respectively, $c(\cdot)$, $a(\cdot)$, and $d(\cdot)$ are known functions specifying the member of the exponential family, and for the survival part

$$p(T_i, \delta_i | \mathbf{b}_i, \boldsymbol{\theta}) = h_i(T_i | \mathcal{H}_i(T_i))^{\delta_i} \exp \left\{ - \int_0^{T_i} h_i(s | \mathcal{H}_i(s)) ds \right\} \quad (7)$$

with $h_i(\cdot)$ given by (2). The integral in the definition of the survival function

$$S_i(t|\mathcal{H}_i(t), \boldsymbol{\omega}_i(t)) = \exp \left\{ - \int_0^t h_0(s) \exp[\boldsymbol{\gamma}^T \boldsymbol{\omega}_i(s) + f(\mathcal{H}_i(s), \mathbf{b}_i, \boldsymbol{\alpha})] ds \right\} \quad (8)$$

does not have a closed-form solution, and thus a numerical method must be employed for its evaluation (e.g. Gauss-Kronrod or Gauss-Legendre quadrature rules). Finally, for the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_{h_0}, \boldsymbol{\alpha})$ standard priors distributions (e.g. independent univariate diffuse normal priors) are used. See [7] for further details.

4 Results

In our analysis the longitudinal outcome \mathbf{y}_i for the i -th patient is given by the time-dependent variable which indicates the cumulative months covered by ACE assumption. In model (1) we fit a linear mixed effect model in which, since the longitudinal trajectories are nonlinear for many patients, we include natural cubic splines in both the fixed and random effects parts in order to allow for flexibility. In model (2) we assume that the risk of death depends on both the current true value of the trajectory and its slope at time t , adjusting according to age, gender and comorbidities covariates at index hospitalization.

It results that all the covariates are associated with the risk for the composite event, except for the current level of the longitudinal profile. In particular, being younger or a female correspond to a higher survival probability, whereas having a higher number of initial comorbidities corresponds to a lower survival probability, as it might be expected. Moreover, the slope of the longitudinal trajectory has a protective role with $\alpha_2 = -0.3985$ (95% CI: $[-0.7508; -0.0059]$), i.e. the higher the slope the higher the survival. Furthermore, we also observe that having a lower slope at the end of the first year leads to larger confidence intervals over time, as shown in Fig. 1, so the uncertainty about the prediction of the survival outcome increases.

5 Conclusion

In healthcare analytics the handling of time-dependent covariates for treatments consumption and adherence are still far from being fully achieved. In this work we observe that the use of the dynamics of drug consumption into a JM framework represents a smart way to include functional data in healthcare analytics and to explore the effects of pharmacological treatments on patients' survival outcomes. However, to obtain a more appropriate functional representation of drug consumption, some improvements may be included into the exposure computation, which is a critical point in pharmacoepidemiology since it can vary by a number of factors [8]. Moreover, a lot of work is needed in order to include simultaneously multiple treatments

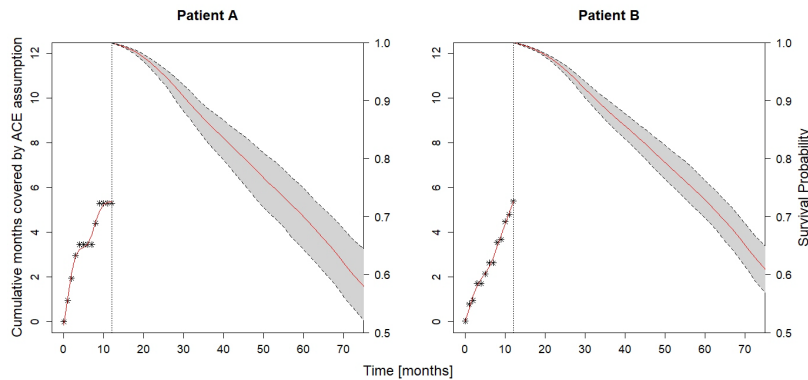


Fig. 1 Longitudinal trajectories (left side of both panels) and survival probability plots (right side) for two male patients with 74 years old and two comorbidities. Moreover, $y_A(12) = 5.293$ and $y_B(12) = 5.392$, $\eta_A(12) = 5.363$ and $\eta_B(12) = 5.370$, $\eta'_A(12) = 0.018$ and $\eta'_B(12) = 0.508$.

in a not trivial way. Nevertheless, since medication nonadherence is usually associated with adverse health conditions and increased economic burden to the healthcare system [9], JM for pharmacoepidemiology represent a highly stimulating area and could also have a strong impact on the society, leading to new answers in the field of policy-making and in the setting of healthcare databases.

References

1. Corrao, G. and Mancina G.: Generating Evidence From Computerized Healthcare Utilization Databases. *Journal of Hypertension*, **65**:490–498, (2015).
2. Rizopoulos, D.: *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman and Hall, (2012).
3. Hickey, G., Philipson, P., Jorgensen, A. and Kolamunnage-Dona R.: Joint Models of Longitudinal and Time-to-Event Data with More Than One Event Time Outcome: A Review. *The International Journal of Biostatistics*, **14**(1), (2018).
4. Mazzali, C. et al.: Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in Lombardy region, 2000 to 2012. In: *BMC Health Service Research*, **16**(234) (2016).
5. Gagne, J.J. et al.: A combined comorbidity score predicted mortality in elderly patients better than existing scores. In: *Journal of Clinical Epidemiology*, **64**(7), 749–59 (2011).
6. Rizopoulos, D.: JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. *Journal of Statistical Software*, **35**(9):1–33, (July 2010).
7. Rizopoulos, D.: The R Package JMBayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC. *Journal of Statistical Software, Articles*, **72**(7):1–46, (2016).
8. Pazzagli, L. et al.: Methods for timevarying exposure related problems in pharmacoepidemiology: An overview. *Pharmacoepidemiology Drug Safety*, **27**(2):148–160, (2018).
9. Karve, S. et al.: Prospective Validation of Eight Different Adherence Measures for Use with Administrative Claims Data among Patients with Schizophrenia. *Value In Health*, **12**(6), (2009).

Bayesian multiscale mixture of Gaussian kernels for density estimation

Stima di densità tramite misture bayesiane multiscala di kernel gaussiani

Marco Stefanucci and Antonio Canale

Abstract In this paper we discuss some preliminary results related to a novel Bayesian nonparametric method for multiscale density estimation. Specifically, we extend the model by [1]—originally developed for compact sample spaces—to deal with data taking values in the whole real line \mathbb{R} . By means of an infinitely-deep binary tree of kernels, we are able to construct a multiscale mixture model able to approximate densities with varying degrees of smoothness and local features. Sampling from the posterior distribution is available with a Markov Chain Monte Carlo method.

Abstract *In questo articolo vengono discussi alcuni risultati preliminari relativi a un nuovo metodo per la stima di densità multiscala nel contesto della statistica bayesiana non parametrica. Nello specifico, il modello introdotto da [1] per spazi campionari compatti, viene esteso per trattare dati definiti su tutta la retta reale \mathbb{R} . Tramite un albero di kernels binario infinitamente profondo è possibile costruire un modello mistura multiscala in grado di approssimare densità con grado di lisciazza variabile e caratteristiche locali. Il calcolo della distribuzione a posteriori è effettuato tramite un algoritmo di catene markoviane Montecarlo.*

Key words: Nonparametric Bayes; Multiscale models; Multiscale stick-breaking

Marco Stefanucci
Università di Padova, Via Cesare Battisti, 241, 35121 Padova e-mail: marco.stefanucci@unipd.it

Antonio Canale
Università di Padova, Via Cesare Battisti, 241, 35121 Padova e-mail: canale@stat.unipd.it

1 Introduction

Nonparametric density estimation is a standard inferential problem. The statistical literature is dominated by single scale methods as, for example, kernel density estimation in frequentist settings ([9]) or Dirichlet process (DP) [3] mixtures (DPM) [2], in the Bayesian context. However, these methods sometimes fail to recover densities with a varying degree of smoothness or local features and a multiscale method may be preferred in these situations. In multiscale analysis, it is not necessary to work with a specific level of resolution, as multcale methods are naturally able to account for different levels of resolution and can adapt to local features. The probably most known multiscale tool are wavelets [10], which became popular in the 90's and are nowadays routinely used in many problems—including density estimation.

We focus here on a Bayesian nonparametric approach ([6]) where the only multiscale alternative to a nonparametric mixture specification is represented by Polya Trees [5]. Polya Trees, however, produce highly spiky density estimates. This unpleasant behaviour can be mitigated using a mixture of Polya Trees [4], but at the cost of more difficult computation. Inspired by the mixture construction of the DPM, [1] proposed a multiscale Bernstein polynomial mixture model, expressing the density f of a continuous and bounded random variable $y \in [0, 1]$ as

$$f(y) = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} \text{Be}(y; h, 2^s - h + 1), \quad (1)$$

where $\text{Be}(a, b)$ is a Beta density with parameter a and b and $\{\pi_{s,h}\}$ is a random set of weights. The double sum in the mixture specification in (1) allows to represent the set of weights and Beta dictionary densities as a binary tree. The parameters of the Beta densities are fixed and directly specified by indexes of the double sum. Hence, a prior for the multiscale mixture (1) can be obtained specifying a stochastic process for $\{\pi_{s,h}\}$ only. The authors introduce such a process inspired by the usual stick-breaking representation of the DP [8] by specifying, for each scale s and node h , two random variables

$$S_{s,h} \sim \text{Be}(1, a), \quad R_{s,h} \sim \text{Be}(b, b), \quad (2)$$

corresponding to the probability of stopping and taking the right path on the binary tree conditionally on not stopping, respectively. Then the weights $\{\pi_{s,h}\}$ are determined by

$$\pi_{s,h} = S_{s,h} \prod_{r < s} (1 - S_{r, g_{shr}}) T_{shr}, \quad (3)$$

where $g_{shr} = \lceil h/2^{s-r} \rceil$ is the node traveled through at scale r on the way to node h at scale s , $T_{shr} = R_{r, g_{shr}}$ if $(r+1, g_{shr+1})$ is the right daughter of the node (r, g_{shr}) and $T_{shr} = 1 - R_{r, g_{shr}}$ if $(r+1, g_{shr+1})$ is the left daughter of the node (r, g_{shr}) .

In [1] an MCMC algorithm for the computation of the posterior distribution is also reported. At first step, data are allocated to a multiscale cluster conditionally on

the current values of the probabilities $\{\pi_{s,h}\}$ while in the second step, conditionally on the cluster allocations, the probabilities are update.

2 A multiscale mixture of Gaussian kernels

Our proposal is inspired by (1), but it is able to deal with data taking values on the whole real line. We assume that the density f is a multiscale mixture of kernels defined on \mathbb{R} ,

$$f(y) = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} K(y; \eta_{s,h}), \quad (4)$$

where $\{\pi_{s,h}\}$ is a set of random weights similarly to [1] and $K(\cdot; \eta)$ is a kernel on \mathbb{R} parametrized by η . Although a general discussion is possible, we focus our attention to the Gaussian kernel, i.e.

$$f(y) = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} N(y; \mu_{s,h}, \sigma_{s,h}^2), \quad (5)$$

where $N(y; \mu, \sigma^2)$ denotes the Gaussian density with mean $\mu \in \Theta_\mu = \mathbb{R}$ and variance $\sigma^2 \in \Theta_{\sigma^2} = \mathbb{R}_+$. The main difference between this model and model (1) is that in our case the parameters of the kernels appearing in (5) are random. This should be though as an advantage, since scale and location parameters of each kernel can adapt to the data. The prior process for the set of weights $\{\pi_{s,h}\}$ is the same as in [1] and is summarized in (2) and (3). Given the unknown nature of $\{\mu_{s,h}\}$ and $\{\sigma_{s,h}^2\}$, we need to define a prior measure for (5) keeping in mind the multiscale nature of the model that we want to retain.

The main idea behind the construction of the process for $\{\mu_{s,h}\}$ is to explore the parameter space Θ_μ in a way such that the values for a given scale (i) have low probability of being near, and (ii) uniformly cover Θ_μ . As for the former point, similar ideas appear in [7]. To this aim, we sample each $\mu_{s,h}$ from a density with a local support that depends on s and h . Specifically, at each scale s , we split Θ_μ in 2^s disjoint intervals and these intervals constitute the supports of the densities which the $\mu_{s,h}$ are sampled from.

The prior process for $\{\sigma_{s,h}^2\}$ relies on the idea that, since kernels on finer scales are related to more local features, the scale parameters are stochastically decreasing.

By using truncated normal distributions for the process regarding $\{\mu_{s,h}\}$ and suitable gamma distributions for the process regarding $\{\sigma_{s,h}^2\}$, the posterior computation is quite easy since we take advantage of the conjugacy, conditionally on the cluster allocation—consistently with [1].

3 Illustration

We illustrate the performance of the procedure using a toy example. We generate a sample of size $n = 100$ from the mixture

$$\pi_1 N(-6, 0.35^2) + \pi_2 N(-2.5, 0.6^2) + \pi_3 N(2, 1.5^2), \quad (6)$$

with $\pi_1 = \pi_2 = \pi_3 = 1/3$. This distribution exhibits three partially overlapping clusters with different scales. The estimation obtained with our approach is compared to a standard DPM of Gaussian kernels with default prior and no hyperprior. The posterior mean densities for both approaches are reported in Figure 1 along with the 95% pointwise credible bands and the true density. Figure 1 shows that the multiscale model can naturally adapt to different degrees of smoothness—here represented by the three different variances of the three normal components—if compared to the standard DPM which, for example, oversmooths the first to the left component. Clearly the performance of the DPM can be improved by using suitable hyper prior distribution on the prior parameters, but this preliminary result sheds light on the importance of adopting a multiscale approach.

A more detailed comparison via simulated and real datasets is subject to ongoing investigations.

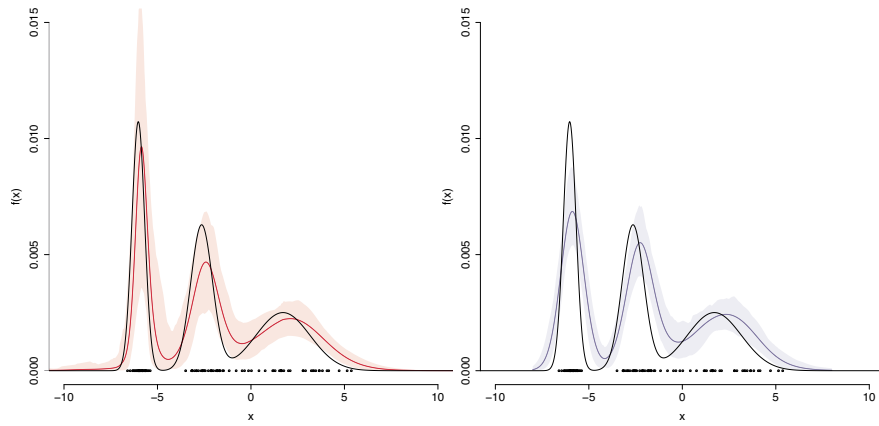


Fig. 1 True density (black lines) corresponding to the mixture of three Gaussian distributions reported in (6) and posterior mean densities computed with our proposal (left, continuous red line) and with a DPM of Gaussians (right, continuous blue line). Shaded areas represent the 95% pointwise posterior credible bands.

Acknowledgements

The authors are supported by the University of Padova under the STARS Grants programme BNP-CD.

References

1. Canale, A. and Dunson, D. B.: Multiscale Bernstein polynomials for densities. *Statistica Sinica* **26**, 1175–1195 (2016)
2. Escobar, M. D. and West, M.: Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588 (1995)
3. Ferguson, T. S.: A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **2**, 209–230 (1973)
4. Hanson, T. E. and Johnson, W. O.: Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association* **97**, 1020–1033 (2002)
5. Lavine, M.: Some aspects of Polya Tree distributions for statistical modelling. *The Annals of Statistics* **20**, 1222–1235 (1992)
6. Mueller, P., Quintana, F. A., Jara, A., Hanson, T.: *Bayesian Nonparametric Data Analysis*. Springer (2015)
7. Petralia, F., Rao, V., and Dunson, D. B.: Repulsive Mixtures. *Advances in Neural Information Processing Systems* **25**, 1889–1897 (2012)
8. Sethuraman, J.: A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650 (1994)
9. Tsybakov, A. B: *Introduction to Nonparametric Estimation*. Springer (2009)
10. Vidakovic, B: *Statistical Modeling by Wavelets*. Wiley (2008)

Dynamic Bayesian clustering of running activities

Clustering Bayesiano dinamico di attività di corsa

Mattia Stival and Mauro Bernardi

Abstract Monitoring sport activities through the use of smart-devices is assuming an increasing importance in several disciplines. Moreover, the ability to store a huge amount of data characterized by a great level of complexity is opening the doors to new challenges and new developments for statistics. In this paper we propose a Bayesian matrix-variate state-space model to cluster the whole trajectories of time series, we introduce a new dataset of running activities recorded by smartwatches, and we propose an exemplary application on such data. Finally, we discuss further developments based on existing literature.

Abstract *Il monitoraggio dell'attività sportiva tramite l'utilizzo di dispositivi intelligenti è di crescente importanza in molteplici discipline e la possibilità di raccogliere un'enorme quantità di dati caratterizzati da una notevole complessità apre le porte a nuove sfide e a nuovi sviluppi nel mondo della statistica. In questo paper proponiamo un modello matrix-variate state-space bayesiano per il clustering di intere traiettorie di serie storiche, utile per il monitoraggio delle attività sportive. Viene inoltre presentato un nuovo dataset di attività di corsa registrate da smart-watch e viene proposta un'applicazione esemplificativa con quei dati. Si discutono infine alcuni sviluppi sulla base della letteratura esistente.*

Key words: State-space models; dynamic clustering; Bayesian methods; smart-watch; sport performance analysis.

Mattia Stival

Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy. e-mail: mattia.stival@phd.unipd.it

Mauro Bernardi

Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy. e-mail: mauro.bernardi@unipd.it

1 Introduction

The evolution of new technologies provides an ever growing amount of data in all the aspects of everyday life and it is rapidly changing the way people make use of information. Athletes of several disciplines, such as running, swimming and cycling, use sport devices that collect geolocalized biometrical and physical data over time. This data are useful for analyzing the performances, in order to check personal physical conditions and to plan future training activities. In this context, data are collected over time as a sequence of activities. Each activity is a high frequency multivariate time series, characterized by complex dependence structures that evolve over time as well as the presence of missing data, outliers, the presence of periodic components, etc. Nevertheless, athletes make extensively use of this kind of data to monitor their performances, which is becoming a relevant topic and can highly benefit the availability of modern statistical methods and models (see, e.g., [1], [4]).

In the context here considered, time series clustering is an interesting statistical approach for analyzing individual activities. In particular, we refer to clustering the whole trajectories of times series aiming to identify groups of activities with similar behaviors. The advantages of performing time series analysis within a state-space framework is threefold: the possibility to consider complex dependencies, such as trend and periodic patterns, the automatic treatment of missing values and the possibility of analyzing data on-line, while they are collected (see, e.g., [3]).

The main contribution of this work is to provide a new Bayesian clustering approach that relies on matrix-variate state-space models, accounting for both cross-sectional and time series dependence among the sequence of multivariate time series that describes the activities. Furthermore, we present an application on real data of 148 running activities of one athlete, coming from a larger dataset composed by more than 2800 running activities recorded by 17 subjects. These data are confidential and were made available for statistical purposes only after a specific request to the users of the on-line platform Strava (www.strava.com), which is extensively used for storing, sharing and analyzing sport data.

For the considered activities, the left panel of Fig.1 shows the time series behavior for the variable Heart Rate in beats per minute (bpm), while the right panel shows the time series behavior for the variable Cumulative Distance in meter (m), for the first 20 minutes of each activity. These two and other variables, such as Latitude, Longitude and Altitude, were measured with a sampling frequency of 1 second. Moreover, for evaluating the enhancement of an athlete during time, it makes sense focusing on the activities of one single subject.

2 The model

Let $y_{p,n,t}$ denotes the observation at time t , for the p -th scalar random variable for activity n , with, $t = 1, 2, \dots, T_n$, $p = 1, 2, \dots, P$ and $n = 1, 2, \dots, N$. Hereafter, T_n denotes the total time in seconds of each activity n , N denotes the number of activities

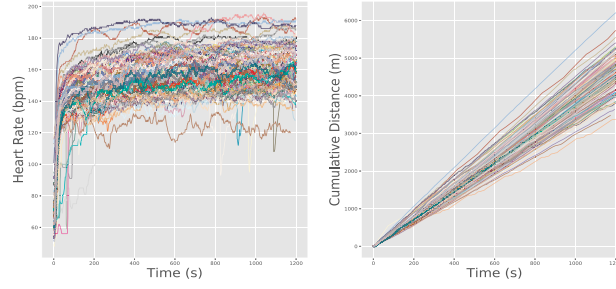


Fig. 1 The behavior over time of Heart Rate (left) and Cumulative Distance (right) of 148 jogging activities of one athlete, measured with a sampling frequency of 1 second. For the two variables, the same colors were used within each activity.

and P denotes the total number of variables measured for each activity. Moreover, we assume that activities can be clustered in one of G different groups depending on the observed variables. We specify the following state-space model for the dynamic evolution of $y_{p,n,t}$:

$$\begin{aligned} y_{p,n,t} &= \mu_{p,t}^{(g)} + \varepsilon_{p,n,t} \\ \mu_{p,t+1}^{(g)} &= \mu_{p,t}^{(g)} + \beta_{p,t}^{(g)} + \eta_{p,t}^{(g)} \\ \beta_{p,t+1}^{(g)} &= \beta_{p,t}^{(g)} + \zeta_{p,t}^{(g)}, \end{aligned}$$

where $\varepsilon_{p,n,t} \sim N(0, \sigma_{\varepsilon,p,n}^2)$, $\eta_{p,t}^{(g)} \sim N(0, \sigma_{\eta,p}^{2,(g)})$, $\zeta_{p,t}^{(g)} \sim N(0, \sigma_{\zeta,p}^{2,(g)})$, with $\varepsilon_{p,n,t} \perp \varepsilon_{p,n,s}$, $\eta_{p,t} \perp \eta_{p,s}$, $\zeta_{p,t} \perp \zeta_{p,s}$, for any $t \neq s$, and $\varepsilon_{p,n,t} \perp \eta_{p,t} \perp \zeta_{p,t}$ for any $n = 1, \dots, N$, $t = 1, \dots, T_n$. To simplify notation, let us denote

$$\mathbf{y}_{n,t} = \begin{bmatrix} y_{1,n,t} \\ y_{2,n,t} \\ \vdots \\ y_{P,n,t} \end{bmatrix}, \quad \boldsymbol{\varepsilon}_{n,t} = \begin{bmatrix} \varepsilon_{1,n,t} \\ \varepsilon_{2,n,t} \\ \vdots \\ \varepsilon_{P,n,t} \end{bmatrix},$$

$$\boldsymbol{\mu}_t^{(g)} = \begin{bmatrix} \mu_{1,t}^{(g)} \\ \mu_{2,t}^{(g)} \\ \vdots \\ \mu_{P,t}^{(g)} \end{bmatrix}, \quad \boldsymbol{\beta}_t^{(g)} = \begin{bmatrix} \beta_{1,t}^{(g)} \\ \beta_{2,t}^{(g)} \\ \vdots \\ \beta_{P,t}^{(g)} \end{bmatrix}, \quad \boldsymbol{\eta}_t^{(g)} = \begin{bmatrix} \eta_{1,t}^{(g)} \\ \eta_{2,t}^{(g)} \\ \vdots \\ \eta_{P,t}^{(g)} \end{bmatrix}, \quad \boldsymbol{\zeta}_t^{(g)} = \begin{bmatrix} \zeta_{1,t}^{(g)} \\ \zeta_{2,t}^{(g)} \\ \vdots \\ \zeta_{P,t}^{(g)} \end{bmatrix},$$

and

$$\mathbf{Y}_t = [\mathbf{y}_{1,t} \ \mathbf{y}_{2,t} \ \dots \ \mathbf{y}_{N,t}], \quad \mathbf{Y}_t = [\boldsymbol{\varepsilon}_{1,t} \ \boldsymbol{\varepsilon}_{2,t} \ \dots \ \boldsymbol{\varepsilon}_{N,t}].$$

The model can be represented in matrix-variate form

$$\begin{aligned} \mathbf{Y}_t &= \sum_{g=1}^G \left(\Lambda^{(g)} \otimes \Theta_t^{(g)} \right) + \mathbf{Y}_t, \\ \alpha_{t+1}^{(g)} &= \mathbf{T} \alpha_t^{(g)} + \psi_t^{(g)}, \end{aligned}$$

where $\Theta_t^{(g)} = \mathbf{Z} \alpha_t^{(g)}$ is the signal with

$$\begin{aligned} \mathbf{Z} &= [\mathbf{Z}_\mu \ \mathbf{Z}_\beta], \quad \mathbf{Z}_\mu = \mathbf{I}_P, \quad \mathbf{Z}_\beta = \mathbf{0}_P, \\ \alpha_t^{(g)} &= \left(\mu_t^{(g)\top}, \beta_t^{(g)\top} \right)^\top, \quad \psi_t^{(g)} = \left(\eta_t^{(g)\top}, \zeta_t^{(g)\top} \right)^\top, \quad \mathbf{T} = \begin{bmatrix} \mathbf{I}_P & \mathbf{I}_P \\ \mathbf{0}_P & \mathbf{I}_P \end{bmatrix}, \\ \Lambda^{(g)} &= (\mathbb{1}(S_1 = g), \mathbb{1}(S_2 = g), \dots, \mathbb{1}(S_N = g)), \end{aligned}$$

where S_n is the mixture indicator for the n -th activity, $n = 1, \dots, N$. Moreover, we assume $\mathbf{Y}_t \sim \text{MVN}(\mathbf{0}, \Sigma_Y \otimes V_Y)$, $\psi_t^{(g)} \sim N(0, \Omega_\psi^{(g)})$ and $\alpha_0^{(g)} \sim N(\mu_0, \sigma_0^2 \mathbf{I}_{2P})$.

The dimensions of vectors and matrices involved in the previous state-space representation are reported in Table 1. A fully conjugate Bayesian approach is adopted to estimate the state-space model parameters.

Table 1 Dimensions of elements in the state-space form of the model.

Matrix/Vector	Dimension	Matrix /Vector	Dimension
\mathbf{Y}_t	$P \times N$	$\alpha_t^{(g)}$	$2P \times 1$
\mathbf{Y}_t	$P \times N$	$\psi_t^{(g)}$	$2P \times 1$
$\Lambda^{(g)}$	$1 \times N$	\mathbf{Z}	$P \times 2P$
$\Theta_t^{(g)}$	$P \times 1$	\mathbf{T}	$2P \times 2P$

3 Results and conclusions

In our application we consider $N = 148$ activities, and we model $P = 2$ variables (Heart Rate and Cumulative Distance), for $T_n = 600$ seconds, for all $n = 1, \dots, N$. The optimal number of groups is $G = 5$, chosen accordingly to DIC (see, e.g., [11]).

Fig. 2 represents the behaviors over time of the extracted posterior signals for the 5 groups, identified with different colors. The visual inspection of the left panel reveals that 2 estimated signals (red and blue) are clearly separated from the remaining three (green, magenta, yellow). This latter group of trajectories is characterized

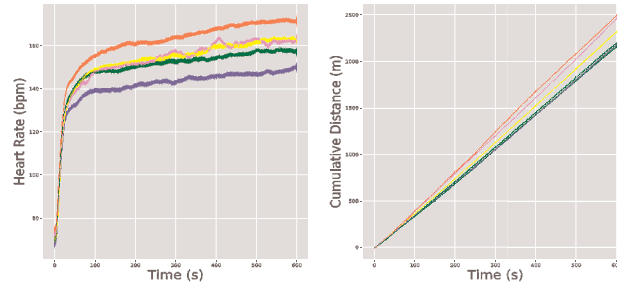


Fig. 2 Posterior draws of signals of Heart Rate and Cumulative Distance for the clusters.

by a similar Heart Rate behavior, which is in contrast to the signals provided by the same clusters on the variable Cumulative Distance. Indeed, focusing on the estimated signals of the joint variables provides a better identification of the behavior of the clusters. For instance, by assuming that the effort of one activity can be described by the behavior of Heart Rate over time, cluster magenta and yellow require similar efforts, but group magenta can be considered better, since the Cumulative Distance covered over time is greater and nearer to the red cluster.

Fig. 3 shows how trajectories are clustered, where the thicker lines are the estimated posterior means of signals for the groups. It is worth noting that, for the red cluster, there exist trajectories far away from their posterior means of signals. These trajectories seem to be a cluster itself but their frequency is low and it seems reasonable the inability of the model on detecting a cluster for these activities.

The exemplary nature of this application should be highlighted. Several issues, such as the choice of variables and the elicitation of prior parameters, have not been discussed. To face these problems, we propose further developments. Within a context of athletes performance measurements several variables are constantly monitored (altitude, temperature, etc.). These variables have been excluded from the model because, despite they can impact the individual assessment, they can introduce bias on the estimated cluster. Their treatment as control variables is a possible solution, and the state-space framework assures the possibility of including them as explanatory variables (see, e.g., [3]). Moreover, a sparsity and parsimonious representation (Σ_T , V_T , $\Omega_{\psi}^{(g)}$) is required to address the problem of sparsity and curse of dimensionality (see, e.g., [12], [2], [7]). The number of individual activities increases day-by-day and, as a further improvement, a Bayesian perspective allows an on-line update of information on the relevant clusters (see, e.g., [8]). Moreover, a matrix variate state formulation allows to capture dependencies due to panel dimension of data (see, e.g., [6]). Assuming that the aim of training is to modify the physical state of one athlete, the number of clusters could be not in principle defined, and new clusters can arise due to enhancements of physical conditions. A

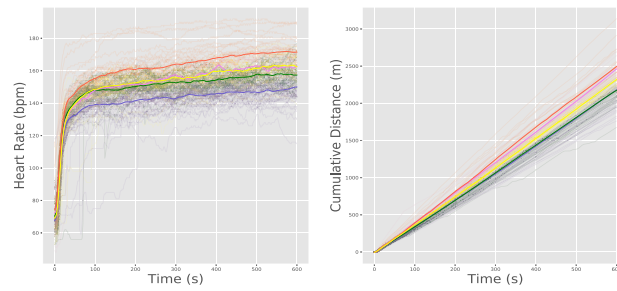


Fig. 3 Time series plot of variables Heart Rate (left panel) and Cumulative Distance (right panel), where each cluster is identified by a different color. The thicker lines are the estimated posterior means of signals for the 5 clusters.

possible solution is related to the adoption of a Bayesian Nonparametric approach for dynamic state-space models (see, e.g., [5], [10], [9]).

References

1. Bartolucci, F., Murphy, T.B.: A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race. *Journal of Quantitative Analysis in Sports*, 11(4):193–203, 2015.
2. Cassese, A., Zhu, W., Guindani, M., Vannucci, M., et al.: A bayesian nonparametric spiked process prior for dynamic model selection. *Bayesian Analysis*, 2018.
3. Durbin, J., Koopman, S.J.: Time series analysis by state space methods, volume 38 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, second edition, 2012.
4. Egidi, L., Gabry, J.: Bayesian hierarchical models for predicting individual performance in soccer. *Journal of Quantitative Analysis in Sports*, 14(3):143–157, 2018.
5. Hjort, N.L., Holmes, C., Müller, M., Walker, S.G.: *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.
6. Huang, L., Bai, J., Ivanescu, A., Harris, T., Maurer, M., Green, P., Zipunnikov, V.: Multilevel Matrix-Variate Analysis and its Application to Accelerometry-Measured Physical Activity in Clinical Populations. *Journal of the American Statistical Association*, pages 1–12, jun 2018.
7. Li, F., Zhang, X.: Bayesian Lasso with neighborhood regression method for Gaussian graphical model. *Acta Math. Appl. Sin. Engl. Ser.*, 33(2):485–496, 2017.
8. Lopes, H.F., Carvalho, C.M.: Online Bayesian learning in dynamic models: an illustrative introduction to particle methods. In *Bayesian theory and applications*, pages 203–228. Oxford Univ. Press, Oxford, 2013.
9. Müller, P., Quintana, F.A., Jara, A., Hanson, T.: *Bayesian nonparametric data analysis*. Springer, 2015.
10. Nieto-Barajas, L., Contreras-Cristán, A.: A Bayesian nonparametric approach for time series clustering. *Bayesian Anal.*, 9(1):147–169, 2014.
11. Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: The deviance information criterion: 12 years on. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):485–493, 2014.
12. Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(3):273–282, 2011.

Employment and fertility in couples: whose employment uncertainty matter most?

Lavoro e fecondità in coppia: il ruolo dell'incertezza lavorativa secondo una prospettiva di genere

Valentina Tocchioni, Daniele Vignoli, Alessandra Mattei, Bruno Arpino¹

Abstract Despite the prevailing belief that economic uncertainty discourages people from having children, empirical findings are mixed. A reason for this inconclusive evidence is that most of previous research has focused on the role of employment uncertainty of men and women in isolation, largely disregarding the employment condition of both partners. We contribute to fill this gap, using data from the United Kingdom Household Longitudinal Survey, and adopting the potential outcome approach to causal inference. We investigate whether and how woman's and/or man's employment uncertainty affects the couple-level propensity to have a first child, and the progression to higher-parity births.

Abstract Nonostante la convinzione prevalente che l'incertezza economica scoraggi dall'avere figli, i risultati empirici sono contrastanti. Fra le motivazioni che giustificano tale eterogeneità, la maggior parte delle ricerche si è concentrata sul ruolo dell'incertezza occupazionale di uomini e donne come soggetti isolati, di fatto ignorando le condizioni lavorative di entrambi i partner. Utilizzando i dati dell'UK Household Longitudinal Survey e adottando un approccio di inferenza causale in termini di risultati potenziali, analizziamo se e come l'incertezza occupazionale femminile e/o maschile incida sulla propensione al livello di coppia ad avere, il primo, secondo o terzo figlio.

Key words: employment uncertainty, fertility, causal inference, UK, panel data, potential outcome approach

¹

Valentina Tocchioni, Università degli Studi di Firenze; email: valentina.tocchioni@unifi.it

Daniele Vignoli, Università degli Studi di Firenze; email: daniele.vignoli@unifi.it

Alessandra Mattei, Università degli Studi di Firenze; email: alessandra.mattei@unifi.it

Bruno Arpino, Università degli Studi di Firenze; email: bruno.arpino@unifi.it

1 Introduction

In recent years, the relationship between employment and fertility has become an increasingly important strand of research, especially in terms of employment uncertainty (Scherer, 2009; Blossfeld & Hofmeister, 2006; Blossfeld, Mills, & Bernardi, 2006). In contemporary Europe, *employment* uncertainty – and the broader concept of *economic* uncertainty – influences both family choices and demographic dynamics, acting as a primary force behind the postponement of childbearing [1]–[3].

Overall, despite the prevailing belief that economic uncertainty discourages people from having children [1], [4]–[6], empirical evidence is mixed (see for example [1], [7]). The sources of these heterogeneous findings are various. Most of previous research has focused on the role of unemployment as an indicator of economic uncertainty, ignoring or downplaying other factors, such as precarious work contracts [8]–[10]. Moreover some studies ignore that unemployment has different meaning and significance depending on the life-course stage and the socio-economic position a person holds [11]. In addition, much of previous research concentrated on women or men in isolation (e.g. [12]), largely disregarding that the labor market conditions of both partners need to be considered in the analysis of how economic uncertainty relates to fertility. In contemporary societies, it is crucial to investigate whether economic uncertainty refers to the man's, the woman's or the couple's employment career. Even more, the type of relationship may play a role in shaping the relationship between economic uncertainty and fertility. When the man is unemployed, married couples may tend to postpone their childbearing, whereas cohabiting couples may react differently, becoming a parent in order to overcome uncertainty in their employment sphere [13]. Finally, economic studies on the economic uncertainty/fertility nexus often conceptualize fertility as a single outcome variable [14], and therefore do not see fertility choices as a succession of transitions in one's life-course [15].

The aim of this work is to advance our understanding of the causal impact of economic uncertainty on fertility. We make three main substantive contributions. First, our analyses account for gender-specific influences of economic uncertainty in explaining couples' childbearing. Second, we anchor our investigation in life-course research. Addressing the economic uncertainty/fertility nexus from the life course perspective means recognizing that family behaviors are intertwined within individuals and over time: fertility does not occur in isolation, but within relationships. Third, we investigate the impact of economic uncertainty not only on the progression to the first child, but also on the progression to higher order births. By adopting a formal framework for causal inference based on the potential outcome approach [16]–[18], we focus on studying the effect of employment uncertainty of one or both partners on inter-couple differences in the propensity to have a(nother) child. More specifically, our analysis addresses the following questions: *Is a couple in which the employment condition of one or both partners is uncertain less prone to have a child than if a couple where both partners have a certain employment condition? Which employment uncertainty matter most – that of the man, of the woman, or both?* We use the potential outcome framework to quantify the impact of employment uncertainty of one or both partners on their birth outcomes up to the third-parity births, and to investigate the heterogeneity of these effects across different familial settings.

We use dyadic data proposing a matching approach under an appropriate strong ignorability assumption. Matching for dyadic data is still rarely employed in family demography research, and will constitute an innovation from a methodological perspective in this context.

2 Data and methods

The study is based on a sample of co-residing couples generated from the first seven waves of Understanding Society, the UK Household Longitudinal Study (UKHLS). In total our observation period spans 2009-2016. We consider a sample of 8,595 co-residing couples interviewed at least for three waves (whose at least two consecutive ones) and where the woman was aged 15-49, had at most two children within that relationship before the first interview date.

Our treatment variable is a multivariate multi-valued variable for the type of economic activity of each member of the couple. Here we consider four types of economic activity: “permanent employment”, “temporary employment”, “unemployment”, and “inactivity”. Therefore, we have 16 different types of treatment statuses according to both man and woman’s activity. Cases are left-censored either at the last interview date or when the woman reaches 50 years old, whichever occurs first.

Our focus is on assessing causal effects on the conception of the first, second and third child. Therefore, we consider three binary outcome variables, equal to 1 for couples who conceive the first, second and third child, respectively. We will conduct three separate analyses for each of the three birth-orders. Under the potential outcome framework, for each outcome variable every couple has 16 potential outcomes, because the couple can conceive or not a child according to each of the 16 treatment values. We will concentrate on the most relevant comparisons, where the condition of both partners working with a permanent employment will be always considered as the reference category. In addition to assessing causal effects for the entire population, we will examine the heterogeneity of the causal effects with respect to specific age groups, the educational level of the partners, and the type of union.

3 Data and causal estimands

As in all longitudinal studies, we observe different individuals for a different number of waves. This implies that we use different sub-samples for estimating birth conceptions each year.

Table 1 reports the number of person-year for type of economic activity of the two partners. The most common condition – which will be considered as the reference category – is that comprising couples where both partners have a permanent employment. Our focus will be on evaluating causal effects defined by comparing potential outcomes for the three outcome variables of interest under the following treatment conditions to potential outcomes under the control level: (1) being a couple formed by an inactive woman and a man with a permanent employment; (2) being a couple where one of the two partners has a temporary employment and the other one has a permanent employment; (3) being a couple formed by an unemployed woman and a man with a permanent employment; and (4) being a couple formed by a woman with a permanent employment and an unemployed man.

Table 1: Person-year over the economic activity of both partners

Economic activity of the woman	<i>Economic activity of the man</i>				<i>Total</i>
	<i>permanent employment (PE)</i>	<i>temporary employment (TE)</i>	<i>unemployed (U)</i>	<i>inactive (I)</i>	
PE	24,233	1,136	790	722	26,881

Tocchioni et al.					
TE	1,452	182	64	64	1,762
U	752	65	392	171	1,380
I	6,281	404	973	936	8,594
Total	32,718	1,787	2,219	1,893	38,617

Source: Wave 1-7 of UKHLS

Table 2 shows the sample sizes according to the type of economic activity, separately for each birth order conception. The sample for the first child's conception is formed by childless couple at the beginning of the observation period (one or both partners may have children from previous relationships); the sample for the second child's conception is formed by couples who had a child within that relationship at the beginning of the observation period; and the sample for the third child's conception is formed by couples who had two children within that relationship at the beginning of the observation period.

Table 2: Child birth's conception and sample size according to the birth order and the economic activity of both partners

a) First child's conception

First child's conception	<i>Economic activity of the woman</i>				<i>Economic activity of the man</i>				Total
	<i>PE</i>	<i>TE</i>	<i>U</i>	<i>I</i>	<i>PE</i>	<i>TE</i>	<i>U</i>	<i>I</i>	
no conception	2,444	184	171	518	2,729	167	197	224	3,317
conception	502	28	25	33	515	28	34	11	588
Total	2,946	212	196	551	3,244	195	231	235	3,905

b) Second child's conception

Second child's conception	<i>Economic activity of the woman</i>				<i>Economic activity of the man</i>				Total
	<i>PE</i>	<i>TE</i>	<i>U</i>	<i>I</i>	<i>PE</i>	<i>TE</i>	<i>U</i>	<i>I</i>	
no conception	1,551	83	83	428	1,816	104	114	111	2,145
conception	460	33	29	169	589	31	46	25	691
Total	2,011	116	112	597	2,405	135	160	136	2,836

c) Third child's conception

Third child's conception	<i>Economic activity of the woman</i>				<i>Economic activity of the man</i>				Total
	<i>PE</i>	<i>TE</i>	<i>U</i>	<i>I</i>	<i>PE</i>	<i>TE</i>	<i>U</i>	<i>I</i>	
no conception	2,037	137	74	661	2,525	124	127	133	2,909
conception	109	6	8	125	205	10	23	10	248
Total	2,146	143	82	786	2,730	134	150	143	3,157

Source: Wave 1-7 of UKHLS

In order to fully understand the relationship between fertility and both partners' economic activity, we will follow up the three samples through the prospective panel. We will investigate the potential gains or losses of births that are attributable to having a combination of both partners' economic activity compared to another one.

References

- [1] M. Kreyenfeld, G. Andersson, and A. Pailhe, "Economic uncertainty and family dynamics in Europe: Introduction," *Demogr. Res.*, vol. 27, no. 28, pp. 835–852, 2012.
- [2] D. Philipov, "Fertility in times of discontinuous societal change: the case of Central and Eastern Europe," Max Planck Institute for Demographic Research, Rostock, 2002.
- [3] H. P. Kohler, F. C. Billari, and J. A. Ortega, "The emergence of lowest-low fertility in Europe during the 1990s," *Popul. Dev. Rev.*, vol. 28, no. 4, pp. 641–680, 2002.
- [4] D. Vignoli, V. Tocchioni, and A. Mattei, "First-Birth Gains and Losses from the First Job in Italy: The Role of Employment Uncertainty," university of Florence, Florence, 2, 2018.
- [5] D. Vignoli, S. Drefahl, and G. De Santis, "Whose job instability affects the likelihood of becoming a parent in Italy? A tale of two partners," *Demogr. Res.*, vol. 26, no. 2, pp. 41–62, 2012.
- [6] D. Vignoli, F. Rinesi, and E. Mussino, "A Home to Plan the First Child? Fertility Intentions and Housing Conditions in Italy," *Popul. Space Place*, vol. 19, no. 1, pp. 60–71, 2013.
- [7] T. Sobotka, V. Skirbekk, and D. Philipov, "Economic recession and fertility in the developed world," *Popul. Dev. Rev.*, vol. 37, no. 2, pp. 267–306, 2011.
- [8] K. Kurz, N. Steinhage, and K. Golsch, "Case study Germany. Global competition, uncertainty and the transition to adulthood," in *Globalization, Uncertainty and Youth in Society*, H.-P. Blossfeld, E. Klijzing, M. Mills, and K. Kurz, Eds. London, UK and New York, US: Routledge, 2005, pp. 51–82.
- [9] C. S. Noguera, T. Castro Martin, and A. S. Bonmati, "The Spanish case. The effect of the globalization process on the transition to adulthood," in *Globalization, Uncertainty and Youth in Society*, H.-P. Blossfeld, E. Klijzing, M. Mills, and K. Kurz, Eds. London, UK and New York, US: Routledge, 2005, pp. 375–402.
- [10] A. C. Liefbroer, "Transition from youth to adulthood in the Netherlands," in *Globalization, Uncertainty and Youth in Society*, H.-P. Blossfeld, E. Klijzing, M. Mills, and K. Kurz, Eds. London, UK and New York, US: Routledge, 2005, pp. 83–104.
- [11] M. Kreyenfeld and G. Andersson, "Socioeconomic differences in the unemployment and fertility nexus: Evidence from Denmark and Germany," *Adv. Life Course Res.*, vol. 21, pp. 59–73, 2014.
- [12] M. Lyons-Amos and I. Schoon, "Differential responses in first birth behaviour to economic recession in the United Kingdom," *J. Biosoc. Sci.*, vol. 50, no. 2, pp. 275–290, 2018.
- [13] H. Inanc, "Unemployment and the timing of parenthood: Implications of partnership status and partner's employment," *Demogr. Res.*, vol. 32, no. 7, pp. 219–250, 2015.
- [14] E. Del Bono, A. Weber, and R. Winter-Ebmer, "Clash of Career and Family: Fertility Decisions after Job Displacement," *J. Eur. Econ. Assoc.*, vol. 10, no. 4, pp. 659–683, Aug. 2012.
- [15] Ø. Kravdal, "The impact of individual and aggregate unemployment on fertility in Norway," *Demogr. Res.*, vol. 6, no. 10, pp. 263–294, Apr. 2002.
- [16] J. Neyman, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Stat. Sci.*, vol. 5, no. 4, pp. 465–480, 1923.
- [17] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Educ. Psychol.*, vol. 66, no. 5, pp. 688–701, 1974.
- [18] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge, US: Cambridge University Press, 2015.

A Functional Data Analysis Approach to Study a Bike Sharing Mobility Network in the City of Milan

Agostino Torti, Alessia Pini and Simone Vantini

Abstract In today's world bike sharing systems are becoming increasingly common in all major cities around the world. To understand the spatio-temporal patterns of how people move by bike through the city of Milan, we apply functional data analysis to study the flows of a bike sharing mobility network. We build a functional-on-functional model taking into account the effects of weather conditions and calendar on the bike flows. In the end, we aim to define a procedure able to predict the future flows between the districts of the city.

Abstract Negli ultimi anni i servizi di bike sharing si sono diffusi capillarmente in tutte le principali metropoli del mondo. Al fine di studiare i pattern spazio temporali di come le persone si spostano nella città di Milano, applichiamo l'analisi dei dati funzionali ai flussi di mobilità di un servizio di bike sharing. Nello specifico, introduciamo un modello funzionale che tenga conto sia della differenza fra i vari giorni della settimana, sia delle condizioni metereologiche. Infine, utilizzando il suddetto modello, ci proponiamo di prevedere i futuri viaggi in bicicletta fra i differenti quartieri di Milano.

Key words: functional data analysis, functional-on-functional model, bike sharing system, Milan.

A. Torti, S. Vantini

MOX- Department of Mathematics, Politecnico di Milano, Pizza Leonardo da Vinci 32, 20133, Milan, Italy.

A. Pini

Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo A. Gemelli 1, 20123, Milan, Italy

A.Torti e-mail: agostino.torti@polimi.it · A.Pini e-mail: alessia.pini@unicat.it · S.Vantini e-mail: simone.vantini@polimi.it

1 Introduction

In the last years, due to urbanization and globalization, cities have been changing at an incredible rate. The tremendous growth of bike sharing systems (BSSs) in all metropolitan areas has urged scientists in developing suitable monitoring and forecasting tools to handle with mobility management and to plan the city of the future [2].

In this context we focus on the city of Milan analysing BikeMi, one of the most famous and older BSS in the city. Our dataset is composed by 42 days from the 25th of January to the 6th of March 2016. The aim of our work is to use bike sharing data to study mobility in Milan providing useful information for the management of urban mobility network. We are interested in understanding the global behaviour of the city and its spatio-temporal patterns: we would like to know how people move by bike, departure and arrival times and venues, studying the variability within and between days. We also aim to quantify how external factors, such as weather conditions (rain, temperature, wind...) or particular events, influence people's mobility behaviour. In the end, we aim to define a procedure able to predict the future bike flows between the districts of the city.

We model BikeMi as a complex network in which the districts of the city are seen as nodes, while bikes moving from one district to another one represent the functional data on each edge. Indeed, since the bike flows from one district to another are continuously dependent on time, we model them making use of tools from Functional Data Analysis (FDA), the branch of statistics dealing with curves, surfaces or anything else varying over a continuum (e.g., [4]). We define a functional flow from district A to district B as a function representing the rate of bikes that are leaving from district A to go towards district B at time t along a single day. Since BikeMi is operative into 39 districts of Milan, our mobility network is composed from $39^2 = 1521$ possible paths. For more details about the network modelling see [6]. To the best of our knowledge, our work is the first one to apply FDA to study the flows of a BSS or any mobility sharing system.

2 Methods and Analysis

To study our functional flows, we develop a pipeline to properly analyse and predict instantaneous flows through a concurrent functional-on-functional model. More precisely, since we are dealing with non negative data, we use a log-linear functional model. The complete pipeline is described in detail in [6], while in this work we briefly report some of the obtained results. For each selected path, the bike flows constitute the dependent functional variables, while the weather conditions and the calendar effects are the known functional and scalar covariates. To make inference on the coefficients of our model, we apply the Interval Wise Testing (IWT) procedure introduced by [3]. The main advantage of this procedure is that it provides an adjusted p-value function which can be thresholded at level α to select the portions

of the domain imputable for the rejection of a null hypothesis (domain selection). To perform model selection, we apply a backward elimination procedure using the adjusted p-value functions as already done by [1] for the analysis of a functional-on-scalar linear model. Successively, after having introduced a novel approach to identify functional influential observations, we define a procedure able to predict future bike flows with the related point-wise empirical prediction bands [6].

We now briefly report some of the obtained results applying the developed approach on the path from Duomo to Duomo, which has deeply been analysed in [6]. First of all, we find out that our flows have a different behaviour during weekdays and weekends, and are affected from both rain and temperature. The IWT procedure allows us to estimate the impact of each covariate on the bike flows at each time of the day and also to evaluate its statistical significance. Firstly, we observe that bike flows during weekends are less than during weekdays and this difference is more evident in the morning: in this time interval the bike trips during weekends are almost 85% less than other days. Second, we find out that bike flows are highly influenced from rain and even a small amount of rain prevents people from renting a bike. However, its effect changes according to the day of the week: during weekdays this reduction is more evident in the morning, while during weekends its effect is almost the same during the whole day. With regards to the temperature, it has a statistically significant impact only between 12:00am-5:00pm: here its coefficient tell us that one degree above (below) the average daily temperature function increases (decreases) bike flows by almost 7%.

Successively, we perform a model diagnostic procedure in a innovative way. We develop an approach able to detect not only functional influential observations, but also able to explain why, by selecting the portions of the domain that bring to identify an observation as an influential one. For example, we find out that days with a high level of rain in a specific time interval influence the estimation of the coefficients of the model in the same time interval.

At this point, our model can be applied to predict bike flows on future days. For example, we show the prediction of a hypothetical weekday fixing the values of rain and temperature. In Figure 1 we plot both the selected weather conditions, on the left, and the predicted curve with the related prediction bands, on the right. We can observe the combined effect of both rain and temperature: the rain, which is present in the morning, implies that there are almost no bike trips in the same time interval; the temperature, which is under its average behaviour, implies that the number of bike trips in the afternoon is less than usual.

3 Conclusions

The aim of our work was to use bike sharing data to study mobility in Milan. To build our analyses, we applied, for the first time, FDA to study the flows of a bike sharing mobility network. We modelled BikeMi like a network with 39 nodes, the districts in which the bike service is operative, analysing the flow data on its 39²

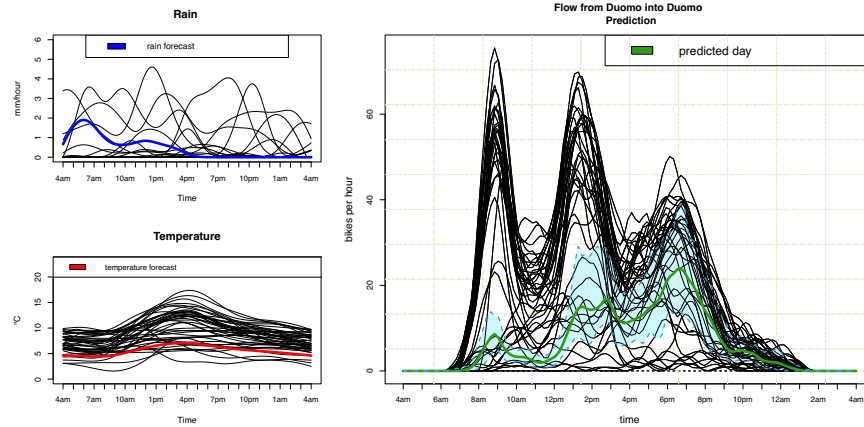


Fig. 1 Left: selected value of rain (in blue) and temperature (in red) compared to the observed weather data (in black) in the period of interest. Right: predicted curve (in green) with the point-wise prediction bands at level 95% (in light blue) compared to the observed flow data (in black) in the period of interest.

available paths, i.e. 1521. Then, we developed a complete pipeline to properly analyse and predict functional data. We applied a functional log-linear model on each path of the network as displayed in the specific case of the flows from Duomo to Duomo. In the end, to dynamically show the results of our procedure, we built an app (created with Shiny [5]), thanks to which it is possible to visualize the results in an interactive way for each selected path. The developed app, whose functioning is deeply presented in [6], allowed us to observe the spatio-temporal patterns of the city, understanding how people move by bike between districts according to the day of the week and the weather conditions. Moreover, this app is accessible through a web link and can be used as an efficient tool to present our work.

In conclusion, our work contributed to implement the study of a bike sharing mobility service in two ways: from a methodological point of view, we defined a procedure to analyse and predict the flows of a bike sharing mobility service through FDA; from an applied point of view, we analysed the bike sharing mobility flows in the city of Milan providing useful information for the mobility management.

References

1. Abramowicz, K., Häger, C. K., Pini, A., Schelin, L., Sjöstedt de Luna, S. and Vantini, S. Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scandinavian Journal of Statistics* (2018)
2. Fishman, E. Bikeshare: a review of recent literature. *Transport Reviews*, 36, 92–113 (2016)
3. Pini, A. and Vantini, S. Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, 29, 407–424 (2017)
4. Ramsay, J.O. and Silverman, B.W. (2005). *Functional data analysis*. Springer, New York.

5. RStudio Inc. shiny: Web Application Framework for R. R package version 0.10.1. <<http://CRAN.R-project.org/package=shiny>> (2014).
6. Torti, A., Pini, A. and Vantini, S. Modeling time-varying mobility flows using function-on-function regression: analysis of a bike sharing system in the city of Milan ,Tech. Rep. 2019 MOX, Politecnico di Milano (2019)

Multiresolution Topological Data Analysis for Robust Activity Tracking

Giovanni Trappolini, Tullia Padellini, and Pierpaolo Brutti

Abstract Multidimensional sensors represent an increasingly popular, yet challenging data source in modern statistics. Using tools from the emerging branch of Topological Data Analysis (TDA), we address two issues frequently encountered when analysing sensor data, namely their (often) high dimension and their sensibility to the reference system. We show how topological invariants provide a tool for detecting change-points which is robust with respect to both the time resolution we consider and the sensor placement.

Abstract *Recentemente, i dati multidimensionali provenienti da sensori hanno guadagnato popolarità, rimanendo tuttavia allo stesso tempo una sfida nella statistica moderna. Usando tecniche e strumenti dall'emergente campo della Topological Data Analysis (TDA), puntiamo ad affrontare due problemi chiave spesso incontrati nell'analisi dei dati provenienti da sensori, ossia la loro elevata dimensionalità e la sensibilità al sistema di riferimento. Mostriamo come gli invarianti topologici offrono uno strumento per rivelare punti di cambio robusti rispetto alla risoluzione considerata e al piazzamento dei sensori stessi.*

Key words: Topological Data Analysis, Sensor Data, High-dimensional Time Series, Multiscale Methods

1 Introduction to Topological Data Analysis

Topology has emerged in the last couple of years as a new tool to analyse highly complex and highly dimensional data, providing a low dimensional and highly interpretable characterization, giving birth to a whole new branch of statistics, Topological Data Analysis (TDA), devoted to recovering topological invariants of the data

Giovanni Trappolini, Tullia Padellini, Pierpaolo Brutti
Sapienza Università di Roma, Piazzale Aldo Moro, 5, 00185 Roma,
e-mail: {giovanni.trappolini, tullia.padellini, pierpaolo.brutti}@uniroma1.it

[8]. The topological or connectivity structure of data in fact can be seen as a statistically meaningful way to represent the “shape of the data”, as topological features of dimension 0 (connected components), in fact represent clusters, while topological features of dimension 1 (loops) can be taught of as periodic relations, and so on. In addition, topological summaries have the advantage of being robust with respect to large classes of deformations and, more critically, are invariant with respect to the coordinate system chosen to represent the data [2].

Since the topology of a point-cloud $\mathbb{X}_n = \{X_1, \dots, X_n\}$ per se is trivial, as it is composed by as many connected components as there are observations and no higher dimensional features, by topology of the data we usually refer to the topology of the support of the generating distribution, \mathcal{X} . The first step in the TDA pipeline is thus to replace the original data with a support estimator such as Devroye’s:

$$\widehat{\mathcal{X}}_\varepsilon = \bigcup_{i=1}^n B(X_i; \varepsilon)$$

where $B(X_i; \varepsilon)$ is the ball of radius ε centered in X_i . While $\widehat{\mathcal{X}}_\varepsilon$ is topologically “richer” than the original point-cloud, it is extremely sensible to the choice of the radius parameter ε . When ε is small, the topology of $\widehat{\mathcal{X}}_\varepsilon$ is close to the one of the point-cloud itself. As ε grows more and more points start to be connected, until eventually, when ε is large, the topology of $\widehat{\mathcal{X}}_\varepsilon$ is that of a point, as shown in Figure 1.

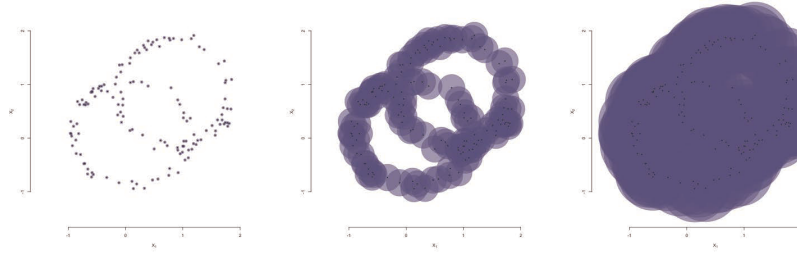


Fig. 1 From left to right: $\widehat{\mathcal{X}}_\varepsilon$ for “small”, “medium” and “large” values of ε .

Rather than choosing an “optimal” ε , the second step in the TDA pipeline is to adopt a multiscale approach and summarizes all resolutions ε into an object called “Persistence Diagram”, a multiset $D = \{(b_i, d_i), i = 1, \dots, m\}$ whose elements are the value of ε corresponding to the creation (b_i) or destruction (d_i) of the i^{th} topological feature.

Points closer to the diagonal represent short-lived features, which may be just noise artefacts, whereas points far away from the diagonal are features that persist at

many different resolutions, hence may be more informative. In addition to being visualization tools, Persistence Diagrams can also be exploited in inferential procedure. Comparison between them can be carried out using the q -Wasserstein distance, which, given two diagrams D and D' is defined as

$$W_q(D, D') = \left[\inf_{\gamma: D \rightarrow D'} \sum_{x \in D} \|x - \gamma(x)\|_\infty^q \right]^{\frac{1}{q}}$$

where the infimum is taken over all bijections $\gamma: D \rightarrow D'$.

2 The Data

We analyse data from REALDISP dataset¹ [1], which is composed of recording from four different types of sensors (recording respectively acceleration, magnetic field orientation, gyro and quaternions), placed on 9 different locations of the body. We consider two different settings, one in which the sensors are positioned by a trained instructor (*Ideal-placement*) and one in which each subject in the analysis is asked to place the sensor on his own (*Self-placement*).

Under this experimental setting, $N = 15$ subject are recorded at a frequency of 50Hz, while performing 30 different every-day activities, ranging from “walking” to “climbing stairs” and “jumping”.

This kind of data fit perfectly in the TDA framework, as the coordinate system is determined by the orientation of the sensor, which depends on how the subject places it (it is enough to twist the sensor a little bit to get different reference systems), and because all the sensors are at least 3-axial (or even 4 in the case of quaternions), meaning that we are dealing with more than 100 variables, which it may not be trivial to jointly account for.

3 Aims and Methods

Our goal is twofold, and consist in exploiting TDA for the following purposes:

1. assess whether a change in activity corresponds to a change in topology.
2. verify that this characterization is robust with respect to how the sensors are placed.

With this in mind, we start with a collection of multivariate time series, in which each subject $i = 1, \dots, N$ is treated as a repetition of the same phenomenon, namely

¹ freely available from <https://archive.ics.uci.edu/ml/datasets/REALDISP+Activity+Recognition+Dataset>

the performance of several physical activities in a prearranged sequence. We take these repetitions separately, and apply to each one of them an overlapping moving windowing procedure. We denote with w the window size, and with s the shift between windows. In each of the windows we estimate the topology of the point cloud that populates it, resulting in a persistence diagram that summarizes the topological information in that particular segments. For each subject we thus obtain a collection of persistence diagrams D_{i1}, \dots, D_{iT} , where T is the number of window of size w , that we exploit to estimate the change in topological structure undergone in the time series. This is achieved through the computation of the wasserstein distance in a "rolling fashion" [5], meaning that each persistence diagram is compared to the previous (and the following) one. In such a way we end up with a time series of distances $W_i = (W_{i2}, \dots, W_{iT})$, that characterize the change in topological structure. Since this procedure was carried out for each "repetition", there will be a collection of such time series distances.

We want to use these repetitions W_1, \dots, W_N , to estimate significative changes in the underlying topological structure. To this end, we fit a local polynomial [7] to the distances W_1, \dots, W_N , and estimate its derivative. We are interested in points in time where this derivative estimate is significantly different from 0, which implies that the topological information is indeed different. In this context we would like these instants to be correlated with a change in physical activity of the subject. Since this analysis may heavily depend on the choice of the length of the window size w , in a scale-space approach [6], we repeat this procedure for different values of it.

Results can be visualized in a SiZer-like [3] fashion, as shown in 3. This map considers for all time instants of interest, on the x-axis, and for all the different windows size used, on the y-axis, the value of the derivative estimates in that point. Each pair (x, y) then, will be encoding a color with the following semantics:

- red: a significant increase in the distance value
- blue: a significant decrease in the distance value
- white: slope not different from 0.

Intuitively, white regions in the plot correspond to time points of topological stability, while colors indicates some sort of change.

4 Results and Discussion

Figure 3 shows how our procedure clearly detect the time point at which the performed activity changes, hence highlighting the potential of topological invariant in the analysis of both high dimensional time series and sensor data. The analysis of multiple widths for the time windows reassures us of the stability of the topological characterization with respect to the sample size, which is critical due to the computational burden naturally related to topological methods. More importantly, the

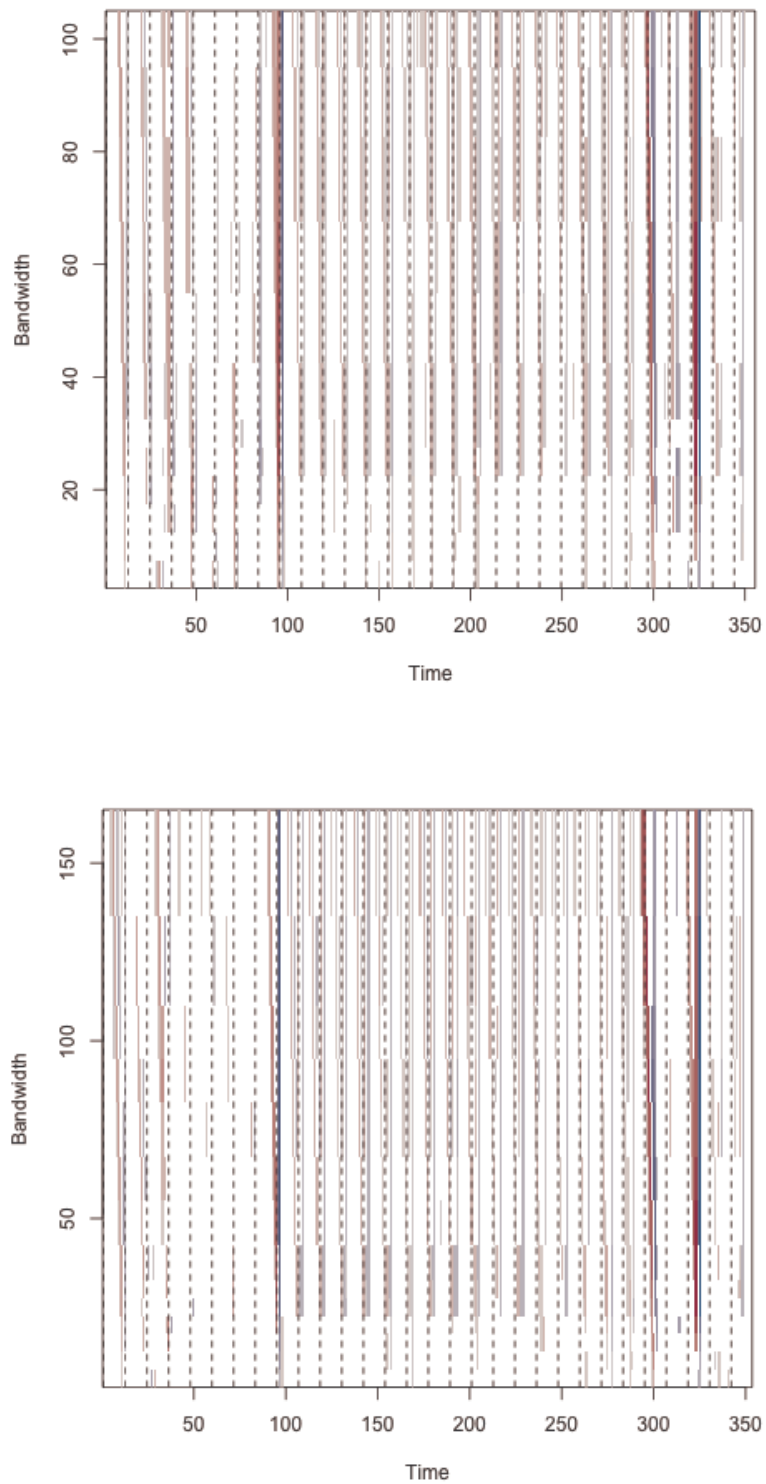


Fig. 2 Derivative maps for Ideal (top) and Self-placement (bottom). Dashed lines correspond to the time points at which the performed activity changes.

results seem to be robust with respect to how the sensor are located, in the sense that there is close to no difference between the Ideal and the Self-placement scenarios.

From an applied perspective, further developments revolve around a proper testing procedure to see whether the two placement scenarios can be considered topologically equivalent and a more thorough analysis of lag sequences s . From a methodological standpoint, we plan to further investigate the theoretical properties of the map we introduced, and its use in changepoint detection.

References

1. Banos, O., Toth, M., Damas, M., Pomares, H., Rojas, I.: Dealing with the effects of sensor displacement in wearable activity recognition. *Sensors* **14**(6), 9995–10,023 (2014)
2. Carlsson, G.: Topology and data. *Bulletin of the American Mathematical Society* **46**(2), 255–308 (2009)
3. Chaudhuri, P., Marron, J.S.: Sizer for exploration of structures in curves. *Journal of the American Statistical Association* **94**(447), 807–823 (1999)
4. Fasy, B.T., Kim, J., Lecci, F., Maria, C., Rouvreau, V.: TDA: Statistical Tools for Topological Data Analysis (2015). URL <https://CRAN.R-project.org/package=TDA>. R package version 1.4.1
5. Gidea, M.: Topological data analysis of critical transitions in financial networks. In: *International Conference and School on Network Science*, pp. 47–59. Springer (2017)
6. Holmström, L.: Scale space methods. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**(2), 150–159 (2010)
7. Loader, C.: *Local regression and likelihood*. Springer Science & Business Media (2006)
8. Wasserman, L.: Topological data analysis. *Annual Review of Statistics and Its Application* **5**, 501–532 (2018)

Semilinear regression trees

Alberi di regressione semilineari

Giulia Vannucci and Anna Gottard

Abstract Tree-based methods refer to a class of predictive models largely employed in many scientific areas. Regression trees partition the variable space into a set of hyper-rectangles, and perform a prediction within each of them. Regression trees are conceptually simple, apparently easy to interpret and capable of dealing with non linearities and interactions. We propose a class of models here called semilinear regression tree, combining a linear component and a tree. These models can handle linear and non linear dependencies and maintains a good predictive performance, while ensuring a simple and intuitive interpretation in a generative model sense. Moreover, we propose an estimation procedure based on evolutionary algorithms.

Abstract *I modelli basati su alberi sono una classe di modelli predittivi ampiamente utilizzate in diverse aree scientifiche. Questi modelli partizionano lo spazio delle covariate in iper-rettangoli, fornendo una previsione in ciascuno di essi. Gli alberi di regressione sono concettualmente semplici, apparentemente di facile interpretazione, in grado di adattarsi a non linearità e interazioni. Si propone una classe di modelli, detti alberi di regressione semilineari, che combinano una componente lineare ed un albero. Questi modelli sono in grado di trattare strutture di dipendenza lineari e non lineari e di mantenere una buona performace predittiva, assicurando un'interpretazione intuitiva in termini di modello generativo. Si propone una procedura di stima dei parametri basata su un algoritmo evolutivo.*

Key words: Regression trees, semilinear regression, evolutionary algorithm

Giulia Vannucci
University of Florence, e-mail: giulia.vannucci@unifi.it

Anna Gottard
DiSIA, University of Florence e-mail: anna.gottard@unifi.it

1 Regression Trees

Regression trees [1] are a class of nonparametric predictive models that leads to a piecewise-constant representation of the regression function. The regression tree model can be written as

$$\mathbb{E}(Y|\mathbf{X}) = T(\mathbf{X}) = T(\mathbf{X}; \mathcal{R}_X, \boldsymbol{\mu}) = \sum_{m=1}^M \mu_{R_m} \mathbb{I}_{(\mathbf{X} \in R_m)} \quad (1)$$

where $\mathcal{R}_X = (R_1, \dots, R_M)$ is a binary recursive partition of $\mathcal{X} : \mathcal{X} = \bigcup_{m=1}^M R_m$, $\mathbb{I}_{(\mathbf{X} \in R_m)}$, $m = 1, \dots, M$, are indicator variables for the hyper-rectangle R_m in \mathcal{R}_X . The parameters μ_{R_m} correspond to $\mathbb{E}(Y|\mathbf{X} \in R_m)$.

One of the most popular algorithm for tree-based regression and classification is *Classification and Regression Trees*, *CART* [1] that recursively finds a binary partition of the explanatory variable space. To select a partition, CART utilizes a top-down greedy approach. The tree is represented in a graph where the nodes denote a splitting of the covariate space and the terminal nodes, or leaves, represent the chose partition. The greedy search algorithm begins from the top of the tree, where all the units belong to the same node, and then repeatedly splits the predictor space into two new nodes. The search is named greedy because at each step the splitting is selected locally, that is conditioned to the previous steps, rather than globally, that is looking ahead and picking a split that conforms with the best final tree.

The advantages of regression trees over many other methods are their ability to include a relatively large number of independent variables and to identify complex interactions among these variables. In addition, regression trees can easily deal with missing data and nonlinear terms. Moreover, the simple graphical representation of the model by a tree graph contributes to its wide use in applications. However, regression trees are not without drawbacks. Along with the high variability of their predictions, the depth of the tree is a tuning parameter of the tree structure that may cause overfitting. They lose interpretability when trees are very large and struggle in modelling steep structures, since they need to perform many splits to recreate a linear dependence [2]. Moreover, most of the algorithms for tree building are based on a greedy recursive partitioning, which is essentially a forward selection of variables. An error in the selection of the variables at the first stage will propagate in all the tree structure.

2 Semilinear regression tree model

For modelling linear and quasi linear dependences at the presence of interactions, we propose a Semilinear regression tree (SRT) as an alternative to ordinary regression trees typically struggling in these situations.

Let $\mathbf{X} = (X_1, \dots, X_p)$ be a vector of explanatory variables and Y be a response variable, and suppose to observe an *iid* sample of n units from this population. We consider modelling $\mathbb{E}(Y|\mathbf{X})$ by $T_{SRT}(\mathbf{X})$, that is

$$\begin{aligned}\mathbb{E}(Y|\mathbf{X}) &= T_{SRT}(\mathbf{X}) = T(\mathbf{X}; \boldsymbol{\beta}, \mathcal{R}_X, \boldsymbol{\mu}) = \beta_1 X_1 + \dots + \beta_p X_p + T(\mathbf{X}; \mathcal{R}_Y, \boldsymbol{\mu}) = \\ &= \sum_{j=1}^p \beta_j X_j + \sum_{m=1}^M \mu_{R_m} \mathbb{I}_{(\mathbf{x}_i \in R_m)}.\end{aligned}\quad (2)$$

The model is characterized by a first part that is a linear component without intercept, with β_1, \dots, β_p unknown parameters, and a second part that is a regression tree, with μ_1, \dots, μ_M unknown parameters. Noticing that a regression tree can be written as a single factor regression model, we consider the proposed model as a general linear model with continuous predictors and a factor with M -levels, that is an ANCOVA in case of Gaussianity assumption. The model (2) in case of p continuous variables and one single factor with 2-levels (one split tree) can be written as

$$\mathbb{E}(Y|\mathbf{X}) = \beta_1 X_1 + \dots + \beta_p X_p + \mu_1 \mathbb{I}_{(X_j \geq s_1)} + \mu_2 \mathbb{I}_{(X_j < s_1)}, \quad (3)$$

where X_j represents the splitting variable and s_1 represents the splitting point. The two indicator functions represent the partitions \mathcal{R} of the tree. An additional split corresponds to the inclusion multiplicatively of a further indicator variable, such as

$$\mathbb{E}(Y|\mathbf{X}) = \beta_1 X_1 + \dots + \beta_p X_p + \mu_1 \mathbb{I}_{(X_j \geq s_1)} + \mu_2 \mathbb{I}_{(X_j < s_1)} \cdot \mathbb{I}_{(X_j \geq s_2)} + \mu_3 \mathbb{I}_{(X_j < s_1)} \cdot \mathbb{I}_{(X_j < s_2)}.$$

This model specification makes it possible to estimate the model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ via the Ordinary Least Square estimator $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

2.1 Evolutionary estimation procedure

As an alternative to the greedy search algorithm that performs a local forward search, we propose an *evolutionary algorithm* able to obtain a globally optimal tree, as proposed in [3].

Evolutionary algorithms are inspired by the principles of genetics and natural selection. In nature, individuals continuously evolving and constantly adapting to their living environment. In Evolutionary algorithms, each individual represents a candidate solution to a problem. At each generation, individuals are evaluated by a fitness function, and the best individuals have a high probability to be selected for reproduction. The operations proposed to the selected individuals are inspired by genetics, like mutation or crossover. More details on evolutionary algorithms can be found in [4].

To infer over the parameters of an SRT model, we propose a procedure based on an evolutionary algorithm, able to estimate both the parameters of the linear component and those of the tree component simultaneously. Firstly (initialization step), the model is initialized to a linear regression model which includes all the X_p covariates and an M -level factor randomly chosen. The parameters are estimated with OLS. Subsequently (perturbation step), the current model is perturbed in the tree part to obtain a new tree. Again the parameters are estimated by OLS. At each step, a comparison between the old and the new models is performed, and the one with minimum MSE is retained (evaluation step). The evolution continues until the last iteration or convergence is reached. At the end of the evolution, the final model is returned.

Let us present the proposed estimation procedure as an example, whose tree component is depicted in Figure 1. The first tree T_1 in the left side is generated randomly in the initialization step with a pre-fixed depth. The random generation of the tree consists of randomly selecting both the splitting variable and the splitting point following the strategy implemented in *Extra-trees* [5] and in *purely* random forests [7].

Starting from the current model

$$E(Y|bmX) = \beta_1 X_1 + \dots + \beta_p X_p + T_1(\mathbf{X}),$$

the tree structure is modified in an evolutionary way, while the linear part of the model is not involved in the modification.

The exploration of the space of the trees is achieved by a perturbation step. The three trees on the right side of Figure 1 sketch the three possibles moves in the perturbation step: *grow*, *prune*, and *mutate*. The move GROW consists in splitting a leaf randomly selected. The move PRUNE consists in the deletion of a node randomly chosen. The move MUTATE allows to change the splitting variable and splitting rule of a node randomly chosen. The idea of a this perturbation step arises from *Bayesian Additive Regression Trees* model [8] and from the *Evtree* algorithm [3].

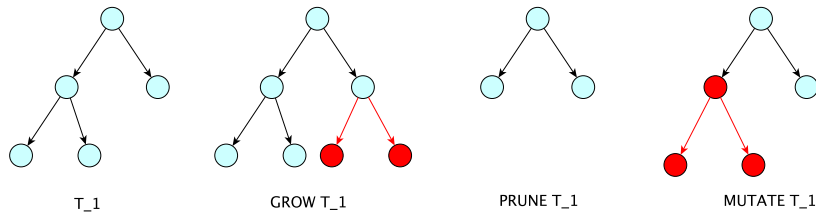


Fig. 1 Example of the perturbation moves of the tree T_1

The evaluation step compares the current and the proposed SRT models in the iterative procedure. As already mentioned, the specification of the model allows the use of the OLS estimator for both β and μ vector of parameters. Once the

parameters are estimated, the goodness of fit of the current and the proposed models is computed. As a measure of goodness of fit we propose to use a MSE with a BIC-type complexity penalty as in [3] and in [9]. The MSE-BIC measure adopted can be specified as

$$\begin{aligned} \text{MSE}_{\text{SRT}} &= n \log(\text{MSE}) + \text{comp}_{\text{SRT}} \\ \text{comp}_{\text{SRT}} &= \lambda \cdot 4 \cdot (G + 1) \cdot \log n \end{aligned} \quad (4)$$

where $G + 1 = M + p + 1$ is the number of estimated parameters, the μ parameters for each of the terminal nodes, the β parameters for the linear part and the residual variance term. The tuning parameter λ is assumed to vary in $[0, 1]$.

The MSE-BIC measure is an alternative to the residual sum of square or R^2 traditionally used in regression model building, which allows to reduce the chance of overfitting.

The recommendation is to keep the depth of the tree as smaller as possible whenever the functional form of the dependence is supposed to be quasi-linear. A short tree can help to avoid overfitting and to improve efficiency. Figure 2 illustrates the possible moves at different depths of the tree when the maximum depth is set to 2.

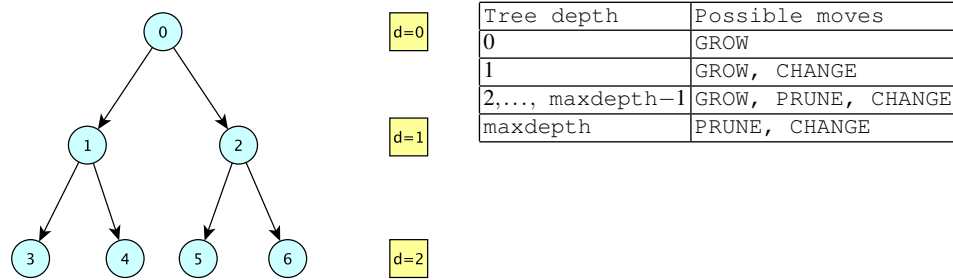


Fig. 2 Possible moves of the perturbation step when the maximum depth of the tree is set to 2.

3 Concluding remarks

Regression trees are interesting models that can easily handle with nonlinear relationship and interactions among variables. Their main advantage is in interpretation, because of the visual diagram that depicts as a tree the partition selected by the algorithm. However, their interpretation is confined in the field of predictive models. On the other hand, regression trees do not boast great predictive performances, particularly in the case of linear or quasi-linear dependence functions.

In this work, we present a class of models particularly useful in these cases of regression trees struggling. We call these models Semilinear Regression trees, con-

sisting of two additive components: a linear part and a tree part. We propose in addition an estimating procedure that looks for the optimal tree component at a global level. Simulation studies, not reported here, showed this algorithm to perform very well when the dependence form does not present strong nonlinearities. As a side, the algorithm provides estimates that can be interpreted as in a generative model. This feature can be considered the main point of strength of Semilinear Regression trees, in the direction of a more interpretable machine learning algorithm than CART.

References

1. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. CRC press (1984).
2. Azzalini, A., Scarpa, B.: Analisi dei dati e data mining. Springer Science & Business Media (2009).
3. Grubinger, T., Zeileis, A., Pfeiffer, K.: evtrees: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R. *Journal of Statistical Software*, 61(1), 1 - 29 (2014).
4. Back, T.: Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms. Oxford University Press, Inc., New York, NY, USA (1996).
5. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. *Machine learning*, 63(1), 3-42 (2006).
6. Breiman, L.: Random forests. *Machine learning*, 45(1), 5-32 (2001).
7. Breiman, L.: Some infinity theory for predictor ensembles. Technical Report 579, Statistics Dept. UCB (2000).
8. Chipman, H. A., George, E. I., McCulloch, R. E.: BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266-298 (2010).
9. Fan, G., Gray, J. B.: Regression tree analysis using TARGET. *Journal of Computational and Graphical Statistics*, 14(1), 206-218 (2005).

A models selection criterion for evaluation of heat wave hazard: a case study of the city of Prato

Un criterio di selezione dei modelli per la valutazione della pericolosità delle ondate di calore: un caso studio della città di Prato

Veronica Villani, Giuliana Barbato, Elvira Romano and Paola Mercogliano

Abstract The main goal of this work is to provide a support for heat waves risk for the city of Prato through the hazard evaluation considering humidex index. The climate analysis has been carried out using a multi-model ensemble of EURO-CORDEX data at high resolution (about 12 km). The approach we propose consists in defining a multicriteria analysis for searching the most appropriate models subset. It is based on the assumption of giving a greater weight to the models with better performance in representing the trends of the variables of interest. After applying the selection criterion, a bias correction method has been used to reduce selected models bias. The analyses have been conducted using the tools available in CLIME service, a horizontal climate service currently developed at CMCC for providing climate data useful for a wide range of users and stakeholders.

Abstract *L'obiettivo principale di questo lavoro è fornire un supporto per il rischio di ondate di calore per la città di Prato attraverso la valutazione della pericolosità climatica considerando l'indice humidex. L'analisi del clima è stata condotta utilizzando un insieme di modelli EURO-CORDEX ad alta risoluzione (circa 12 km). L'approccio che proponiamo consiste nel definire un'analisi multicriterio per la ricerca del sottoinsieme di modelli più appropriato. Si basa sul presupposto di dare un peso maggiore ai modelli con prestazioni migliori nel rappresentare gli*

Veronica Villani

REgional Models and geo-Hydrological Impacts, Centro Euro-Mediterraneo sui Cambiamenti Climatici, Italy; e-mail: veronica.villani@cmcc.it

Giuliana Barbato

REgional Models and geo-Hydrological Impacts, Centro Euro-Mediterraneo sui Cambiamenti Climatici, Italy; e-mail: giuliana.barbato@cmcc.it

Elvira Romano

Università degli Studi della Campania Luigi Vanvitelli, Dipartimento di Matematica e Fisica, Italy; e-mail: elvira.romano@unicampania.it

Paola Mercogliano

CIRA Italian Aerospace Research Centre, Meteorology Laboratory, via Maiorise, Capua (CE), Italy; e-mail: p.mercogliano@cira.it

andamenti delle variabili di interesse. Dopo aver applicato il criterio di selezione, è stato utilizzato un metodo di correzione del bias per ridurre il bias dei modelli selezionati. Le analisi sono state condotte utilizzando gli strumenti disponibili nel servizio CLIME, un servizio climatico orizzontale attualmente sviluppato presso il CMCC per la fornitura di dati climatici utili a una vasta gamma di utenti e stakeholders.

Key words: climate changes, impact studies, heat waves, high-resolution climate projections, multi-model ensemble, multicriteria approach, bias correction.

1 Introduction

The purpose of this work is to study the variation of heat waves for the city of Prato. The parameterization of the heat wave hazard has been performed through the humidex index that was used as the climate parameter for evaluation of the thermal discomfort due to high temperatures (heat waves). Heat waves are defined as “an extended period of unusually high atmosphere-related heat stress, which causes temporary modifications in lifestyle and which may have adverse health consequences for the affected population [1]”. The magnitude of heat wave may be aggravated by the presence of urban settlements: phenomenon known as Urban Heat Island. Heat waves could have many negative impacts, including excess negative health outcomes, energy (consumption and peak demand) and water consumption. Climate analysis was performed using CORDEX regional climate model (RCM) simulations available over the European domain (EURO-CORDEX) at resolution 0.11 degree (about 12 km) forced by different global climate models. EURO-CORDEX is the European branch of the international CORDEX initiative, which is a program sponsored by the World Climate Research Program (WRCP, <http://wcrp-cordex.ipsl.jussieu.fr>) to organize an internationally coordinated framework to produce improved regional climate change projections for all land regions world-wide. The CORDEX results will serve as input for climate change impact and adaptation studies within the timeline of the Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC) and beyond. More information on the EURO-CORDEX initiative are available at the following link <http://www.euro-cordex.net>. The simulations taken into account are obtained according to the IPCC, RCP4.5 and RCP8.5 scenarios. RCPs are time and space dependent trajectories of concentrations of greenhouse gases and pollutants resulting from human activities, including changes in land use. RCPs provide a quantitative description of concentrations of the climate change pollutants in the atmosphere over time, as well as their radiative forcing in 2100 (specifically, RCP4.5 achieves an overall impact of 4.5 watts per square meter by 2100 and RCP8.5 of 8.5 watts per square meter) [2]. The climate anomalies were evaluated over the future periods 2021-2050 and 2071-2100 with respect to the control period 1981-2010. Two climate variables have been taken into account: maximum temperature (°C) and relative humidity (%). More-

over, the climate projections over the city of Prato have been elaborated using a multi-model ensemble approach. To take into account the fact that not all models have the same performances, the results in terms of average (ensemble mean), have been obtained selecting only the most appropriate models using a multicriteria analysis. After applying the selection criterion, a bias correction method has been used to reduce selected models bias. The climate analysis was performed through specific tools developed in CLIME service, an interactive multi-user platform developed by Regional Models and geo-Hydrological Impacts (REMHI) division of CMCC Foundation (Euro-Mediterranean Center on Climate Change), within the CLARA project funded by European Commission, Horizon 2020.

2 Methodology

The heat wave hazard for the city of Prato was performed considering humidex index. The humidex index is a climate parameter adopted for evaluation of the thermal discomfort due to high temperatures based on daily maximum temperature and relative humidity. The humidex index used by the Canadian meteorological service is defined by the following equation [3]:

$$I_H = T_{air} + 0.5555[6.11e^{5417.7530(\frac{1}{273.16} - \frac{1}{T_{dew}})} - 10] \quad (1)$$

in which T_{air} is the air temperature and T_{dew} is the dew point temperature. The latter is a function of the relative humidity H_R of the air, as shown in the Magnus-Tetens formulation:

$$T_{dew} = \frac{b_T \theta(T, H_R)}{a_T - \theta(T, H_R)} \quad (2)$$

where:

$$\theta(T, H_R) = \frac{a_T T}{b_T + T} + \ln(H_R). \quad (3)$$

This index is a measure of perceived heat that results from the combined effect of excessive humidity and high temperature. Five categories are introduced to define humidex index (as shown in Table 1). The main goal of this work is to provide

Table 1 Range of humidex index (Degree of comfort)

Category	Index value	Description
1	< 20	Not classified
2	20-30	Little discomfort
3	30-40	Some discomfort
4	40-45	Great discomfort; avoid exertion
5	> 45	Dangerous; Heat stroke possible

a support for heat wave risk for the city of Prato. The climate analysis has been carried out using a multi-model ensemble of EURO-CORDEX data, as done in several recent literature works [4], [5]. The climate simulations used in this work consists of eighteen GCM-RCM combinations carried out in the frame of EURO-CORDEX. In fact, if we consider only a single RCM-GCM combination, we obtain only one of many potential outcomes. To sample the range of possible outcomes, and uncertainty associated with particular RCMs and/or GCMs, it is necessary to provide ensemble simulations combining different RCMs with different GCMs, as it is done within the CORDEX framework (further information are available in the report Guidance for EURO-CORDEX climate projections data use published by the EURO-CORDEX community). The sources of uncertainties can be grouped into three major categories: scenario uncertainty, internal climate variability and model uncertainty. To take into account the fact that the models do not have the same performances, the results in terms of average (ensemble mean) have been obtained using a multicriteria analysis that consists in searching the most appropriate models subset. In literature, different experimental approaches for the selection of models to be included in the ensemble are available. According to some approaches, known as past-performance approaches (e.g. [6]; [7]), climate models are often selected based on their skill in representing, for the present climate, the trends of the variables of interest (medium and extreme values). Other kind of approaches, known as envelope approaches (e.g. [8]; [9]; [10]; [11]; [12]), aim at reducing the number of models to be included in the ensemble allowing to represent a wide range of possible future scenarios. Depending on the climate change impact assessment, different types of variables and extreme indicators are considered as input of the envelope approaches. Both the methods mentioned above have some limitations. Specifically, the past-performance approaches consider only the abilities of the models in representing the present climate by omitting the characteristics of the future scenarios, while the envelope approaches consider only the convergence of the climatic anomalies of the single models ignoring the performance on the present. To overcome this limitations, in literature several approaches have been presented to select climate models combining the envelope approach and the past-performance approach ([13]; [14]). The multicriteria approach proposed in this work combines the different experimental approaches in a three sequential steps analysis by extending the procedure of [15]. The goal is to select an appropriate subset of models which still represents all possible simulated futures in terms of maximum temperature and relative humidity and in terms of future changes in the climatic extreme index considered. In addition, the aim is to select only models that have sufficient skill in simulating the present day climate in the study area. After applying the criterion mentioned above, the climatic data of the selected simulations have been corrected in order to reduce their systematic error through non-parametric quantile mapping using robust empirical quantiles ([16]; [17]).

3 Preliminary results

Some preliminary results obtained through the methodology described above are shown below. To assess the statistical significance of the trend of Humidex indicator calculated through EURO-CORDEX models for the city of Prato, the Mann-Kendall test was used with a 95% confidence level. Figures show the annual time series of Humidex indicator with the trend values, considering scenarios RCP4.5 and RCP8.5: the statistically significant trends are identified by an asterisk. The shaded areas represent the range of the uncertainty of the projections. Furthermore, the black line represents the observations of the city of Prato for the period 1996-2015. Figure 2 shows results in agreement with those of Figure 1 considering only the models subset obtained through multicriteria approach. In this case the range of uncertainty associated to the ensemble is reduced in width. This means that our proposal is able to find a trade off between ensuring all possible simulated futures and selecting models that have sufficient skill in simulating the present day climate in the studied area. Moreover, in Figure 3, it can be observed that the correction of the systematic error was successful.

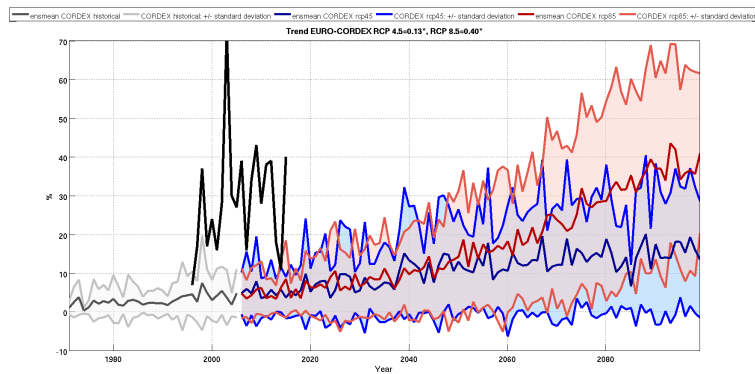


Fig. 1 Climate projections of the EURO-CORDEX models for Humidex indicator, considering the historical period (in gray) and the RCP4.5 (in blue) and RCP8.5 (in red) scenarios. The dark thick line indicates the mean climatic projection (ensemble mean), calculated by averaging the values of all the simulations considered; the shaded areas represent the range of the uncertainty of the projections. All EURO-CORDEX simulations currently available for the variables used were considered. Furthermore, the black line represents the observations of the city of Prato for the period 1996-2015.

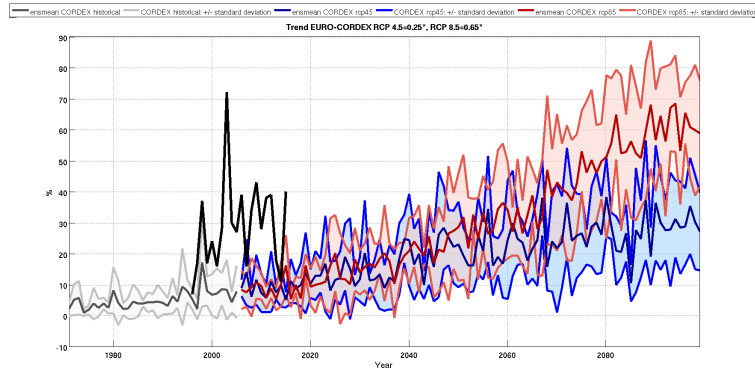


Fig. 2 Climate projections of the EURO-CORDEX models for Humidex indicator, considering the historical period (in gray) and the RCP4.5 (in blue) and RCP8.5 (in red) scenarios. The dark thick line indicates the mean climatic projection (ensemble mean), calculated by averaging the values of all the simulations considered; the shaded areas represent the range of the uncertainty of the projections. EURO-CORDEX simulations subset obtained through multicriteria approach proposed in this work is considered. Furthermore, the black line represents the observations of the city of Prato for the period 1996-2015.

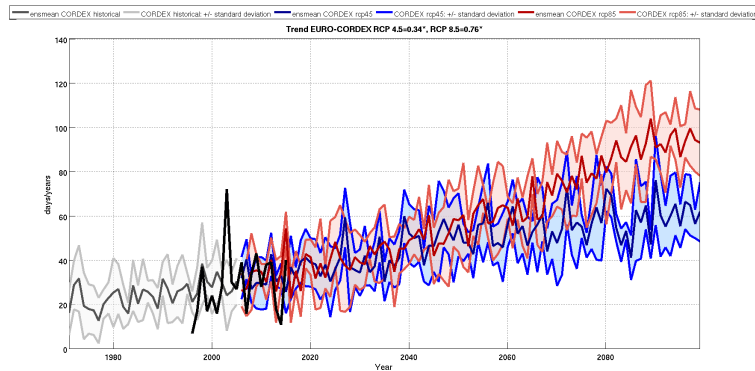


Fig. 3 The bias corrected climate projections of the EURO-CORDEX models for Humidex indicator, considering the historical period (in gray) and the RCP4.5 (in blue) and RCP8.5 (in red) scenarios. The dark thick line indicates the mean climatic projection (ensemble mean), calculated by averaging the values of all the bias corrected simulations considered; the shaded areas represent the range of the uncertainty of the projections. The bias corrected EURO-CORDEX simulations subset obtained through multicriteria approach proposed in this work is considered. Furthermore, the black line represents the observations of the city of Prato for the period 1996-2015.

References

1. Robinson P.J., 2001. *On the Definition of a Heat Wave*. Journal of Applied Meteorology 40(4): 762–75. WOS:000168022100007.

2. Moss R.H., Edmonds J.A., Hibbard K.A., Manning M.R., Rose S.K., van Vuuren D.P., Carter T.R., Emori S., Kainuma M., Kram T., Meehl G.A., Mitchell J.F.B., Nakicenovic N., Riahi K., Smith S.J., Stouffer R.J., Thomson A.M., Weyant J.P., Wilbanks T.J., 2010. *The next generation of scenarios for climate change research and assessment*. Nature 463:747–756. DOI:10.1038/nature08823.
3. Masterson J., Richardson F.A., 1979. *Humidex, a Method of Quantifying Human Discomfort Due to Excessive Heat and Humidity*. Downs view, Ontario: Environment Canada.
4. Jacob D., Petersen J., Eggert B., et al, 2014. *EURO-CORDEX: new high-resolution climate change projections for European impact research*. Reg Environ Chang 14:563–578. DOI: 10.1007/s10113-013-0499-2.
5. Kotlarski S., Keuler K., Christensen O.B., et al, 2014. *Regional climate modeling on European scales: A joint standard evaluation of the EURO-CORDEX RCM ensemble*. Geosci Model Dev 7:1297–1333. DOI: 10.5194/gmd-7-1297-2014.
6. Pierce D.W., Barnett T.P., Santer B.D., Gleckler P.J., 2009. *Selecting global climate models for regional climate change studies*. Proc. Natl. Acad. Sci. U. S. A. 106(21): 8441–8446, doi: 10.1073/pnas.0900094106.
7. Biemans H., Speelman L.H., Ludwig F., Moors E.J., Wiltshire A.J., Kumar P., Gerten D., Kabat P., 2013. *Selecting global climate models for regional climate change studies*. Proc. Natl. Acad. Sci. U. S. A. 106(21): 8441–8446, doi: 10.1073/pnas.0900094106.
8. Cannon A.J., 2014. *Selecting GCM scenarios that span the range of changes in a multi-model ensemble: application to CMIP5 climate extremes indices*. J. Clim. 28: 1260–1267, DOI: 10.1175/JCLI-D-14-00636.1.
9. Houle D., Bouffard A., Duchesne L., Logan T., Harvey R., 2012. *Projections of future soil temperature and water content for three Southern Quebec forested sites*. J. Clim. 25(21): 7690–7701, DOI: 10.1175/JCLI-D-11-00440.1.
10. Immerzeel W.W., Pellicciotti F., Bierkens M.F.P., 2013. *Rising river flows throughout the twenty-first century in two Himalayan glacierized watersheds*. Nat. Geosci. 6(8): 1–4, DOI: 10.1038/ngeo1896.
11. Sorg A., Huss M., Rohrer M., Stoffel M., 2014. *The days of plenty might soon be over in glacierized Central Asian catchments*. Environ. Res. Lett. 9(10), DOI: 10.1088/1748-9326/9/10/104018.
12. Warszawski L., Frieler K., Huber V., Piontek F., Serdeczny O., Schewe J., 2014. *The inter-sectoral impact model intercomparison project (ISI-MIP): project framework*. Proc. Natl. Acad. Sci. U. S. A. 111(9): 3228–3232, DOI: 10.1073/pnas.1312330110.
13. Giorgi F., Mearns L.O., 2002. *Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the “reliability ensemble averaging” (REA) method*. J. Clim. 15(10): 1141–1158.
14. Rianna G., Reder A., Villani V., Mercogliano P., 2017. *Variations in Landslide Frequency Due to Climate Changes Through High Resolution Euro-CORDEX Ensemble*. In: Mikos M., Casagli N., Yin Y., Sassa K. (eds) *Advancing Culture of Living with Landslides*. WLF 2017. Springer, Cham; pp 237–242 - DOI: 10.1007/978-3-319-53485-5-27.
15. Lutz A. F., ter Maat H.W., Biemans H., Shrestha A.B., Westerd F., Immerzeel W.W., 2016. *Selecting representative climate models for climate change impact studies: an advanced envelope-based selection approach*. Int. J. Climatol. 36: 3988–4005. DOI: 10.1002/joc.4608.
16. Villani V., Rianna G., Mercogliano P., Zollo A.L., 2015. *Statistical approaches versus weather generator to downscale RCM outputs to slope scale for stability assessment: a comparison of performances*. Electronic Journal of Geotechnical Engineering Vol.20.4 pp 1495–1515.
17. Gudmundsson L., Bremnes J. B., Haugen J. E., Engen Skaugen T., 2012. *Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations - a comparison of methods*. Hydrology and Earth System Sciences 9, 6185–6201.

Digital Inequalities and ICT Devices: The ambiguous Role of Smartphones

Laura Zannella, Marina Zannella

Abstract

This study builds on micro-data from the community European survey on “the ICT usage in household and by individuals” to analyze the digital divide in Italy based on three different dimensions: the technical means (kind of broadband connection and type of device), the use of internet (kind of activities) and the digital skills. We analyze changes occurred over the 2010-2016 period in the use of internet of Italians, using multivariate statistical analysis, we investigate whether there are significant differences among individuals, and particularly among different generations and socio-economic groups, in the use of specific devices such as smartphone or laptop to access internet. Our particular aim is to disentangle the relation between the use of the smartphone and the development of digital skills. Our findings highlight that the solely use of the smartphone is more common among segments of the population characterized by a first level digital divide. Moreover, the results document a negative relation between the exclusive use of the smartphone to access internet and individual digital skills.

Key words: ICT; Digital divide; Device; Smartphone; Generations

Laura Zannella
Istituto Nazionale di Statistica (Istat)

Marina Zannella
Sapienza Università degli studi di Roma

1 Introduction

The notion of digital divide has changed over time at the same pace as new information and communications technologies (ICT) evolved. As a consequence of this characteristic, the digital divide has been described as a mobile target continuously repositioning itself (Compaine 2001). At the beginning of the 1990s, the digital divide was defined through a dichotomous approach based on the access/non-access to new technologies. However, together with the gradual diffusion of internet, a paradigm shift occurred in the scientific debate: the digital divide was no longer defined as a one-dimensional concept but as a multidimensional one including, in addition to traditional measures of access, also information on the degree of autonomy in the use of ICT, the kind of use (e.g. activities performed) and the skills acquired (Hargittai and Di Maggio 2001). Such paradigm shift has led to a new interpretation of the digital divide as digital inequalities. For instance, the widespread of ICT has created new possibilities of being connected, in any place and moment, through a plurality of devices. The determining factor in the measurement of the digital divide has therefore shifted from the possibility of accessing the network to having the appropriate digital skills to be able to make full use of the opportunities offered by ICT.

In 2016, Italy was still characterized by a first-level digital divide; different opportunities to access ICT were largely explained by generational, social and territorial factors. In most European countries the regular use of internet reaches 90% for the population up to the age of 54, while in Italy levels close to saturation are recorded only for the "network generation" (90.3%). In the last decade, the number of users has almost doubled and the generation that has proved to be the most dynamic is that of the "transition" (individuals born between 1966-1976) which has recorded increases in the ICT usage such as to reduce the gap with the generations grown in the digital era. Nevertheless, despite the availability of different types of devices to connect, in 2016 only 7.5% of individuals aged 15 and over report to use multiple devices and the most widespread tool for surfing the net is the smartphone.

The survey on ICT usage in households and by individuals conducted by Istat provides harmonized information and indicators at the European level in order to evaluate, monitor and redirect the policies implemented by the government in the field of ICT¹. The need to collect harmonized data on the use of ICT derives from the objectives set by the Council of Europe in 2000. In March 2010 the European Commission launched the Europe 2020 strategy for smart, sustainable and inclusive growth. The Digital Agenda for Europe is one of the seven flagship initiatives of the Europe2020 strategy and aims to establish the key role of information and communication technologies for social and economic development.

¹ The survey, which is cofounded by the European Community, conforms to the regulation (EC) No. 808/2004 of the European Parliament and of the Council of 21 April 2004 concerning Community statistics on the information society.

This paper focuses on Italy to investigate changes in the use of internet and digital inequalities over the last decade. In particular, we rely on micro-data from the Survey on the Use of New Technologies by Families and Individuals (Istat) for the years 2006 and 2016, to answer the following research questions: Do greater opportunities of being connected reduce digital inequalities? Is the use of specific devices to access internet (e.g. smartphones, PCs, tablets) characterized by generational and cultural specificities? Can we consider the smartphone as a driver of digital inclusion?

2 Data and methods

The study builds on micro data from two different years (2006 and 2016) of the Italian Survey on information and communication technology (ICT) in household and by individuals. The survey, which is carried out by Istat on annual basis within the broader “Aspects of daily life” survey, is part of the European Community statistics on the information society establishing the legal basis for harmonized statistics on ICT usage in household and by individuals (Commission Regulation No 808/2004).

Given the evolving situation of information and communication technologies, the survey’s model questionnaire changes every year. The principal aim of this survey is to supply users with a wide range of indicators on information society including those related to: Internet activities and connection used (fixed and mobile broadband), e-Commerce, digital skills, e-skills; the mobile use of the Internet - ubiquitous connectivity; cloud computing, e-government; trust and security. ICT surveys are also one of the major sources of data for the Digital Agenda Scoreboard measuring progress of the European digital economy. Information comprises, in addition to the socio-demographic characteristics of individuals and households, the following aspects: access to selected IC technologies, use of computers, location, frequency of use, activities, use of the Internet, internet, e-skills, the mobile use of the Internet. The survey represents therefore a precious source of information for studying relevant aspects of the ICT diffusion and usage as the main socio-economic determinants of the digital divide and digital inequalities in our country.

A first step of this study consists in analyzing changes occurred between 2006 and 2016 in the first-level digital divide. Then, we focus on the most recent data available for year 2016 to analyze digital skills. Digital skills are identified in line with the Digital Competence Framework developed by the European Commission in collaboration with national statistical offices. The framework identifies five competence domains: information, communication, content creation, safety and problem solving. Information about activities realized by the respondents during the previous three months in four of these domains are collected by the survey. Thus, we select a set of activities that reflect the competences outlined within each domain of the framework (with a minimum of 4 and a maximum of 7 activities selected). For

example, for digital skills in the field of information, five activities have been selected: copying or moving a file or folder, obtaining information from Public Administration websites, finding information on goods and services, finding health information. Hence, based on the information on the activity performed, for each respondent and for each of domain the following three levels of digital skills are identified: none, basic and above basic. Once these three levels of skills are computed for each of the four dimensions, an overall composite indicator is computed following a similar logical approach.

Our particular aim is to investigate whether and to what extent the use of specific devices associates with digital inequalities related to both first and second levels digital divides. A particular attention will be given to the use of the smartphones. Our specific aim is to evaluate if the use of the smartphone can be considered a driving force in the adoption of new technologies and, thus, reduce the digital inequalities among individuals and households. Taking into account the hierarchical structure of the data (territory, individuals), the analysis is carried out using a multilevel approach. Our dependent variable is the rate of utilization of smartphone by regular internet regular Internet users. The first level units consist in the area of residence of the individuals, information in this regard comprise: demographic amplitude, metropolitan area, region classified according to the degree of development, etc. Second level units are individuals for which we include regarding the main socio-demographic characteristics such as sex, age, educational qualification, and employment status.

3 Descriptive findings

Descriptive findings provide a first empirical evidence of the relationships existing between the utilization rates of the different devices by regular Internet users and some of their "structural" characteristics such as gender, age, level of education, professional status, position in the profession, frequency of use of the Internet. As shown in Fig.1, the choice of the device used to surf is characterized by generational specificities. Young people up to the age of 34 are more likely to combine the use of PC and mobile phone, although almost a quarter of individuals in this age group exclusively use the smartphone to access internet. By contrast, the exclusive use of the PC prevails among population aged 55 and over. The use of multiple devices is more widespread among men, as well as the exclusive use of the pc (which is particularly widespread after age 55). Women prefer, instead, the exclusive use of the mobile phone.

Figure 1 – Individuals regularly using the internet by kind of device and age group (percentage values). Year 2016

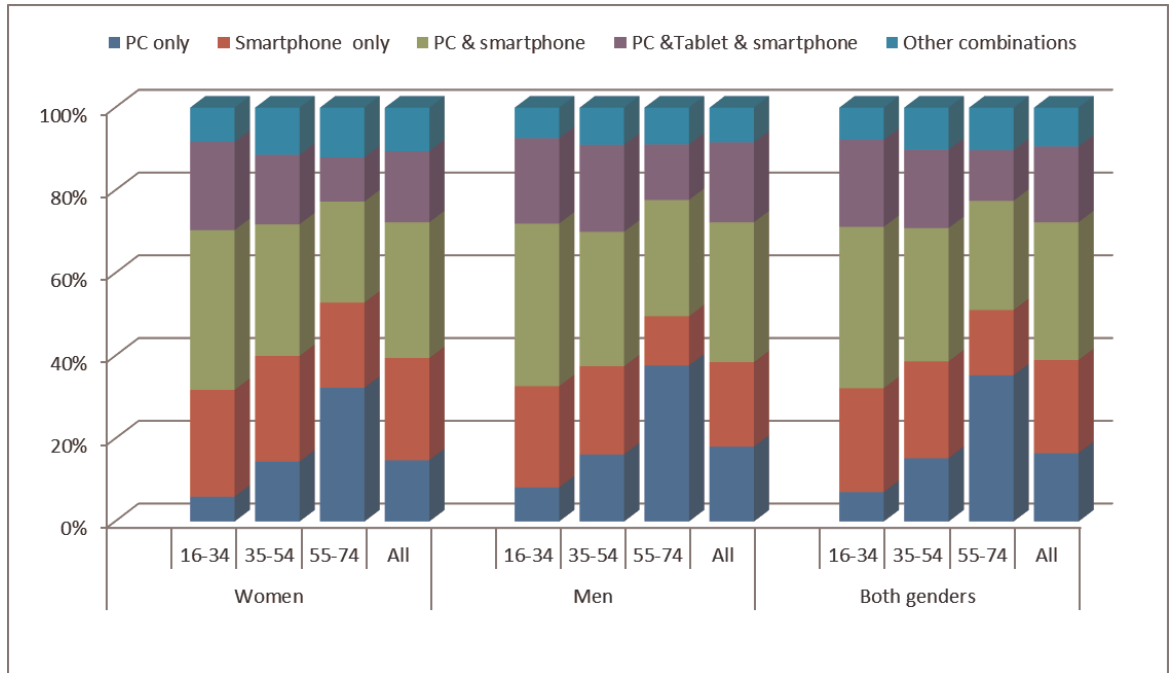
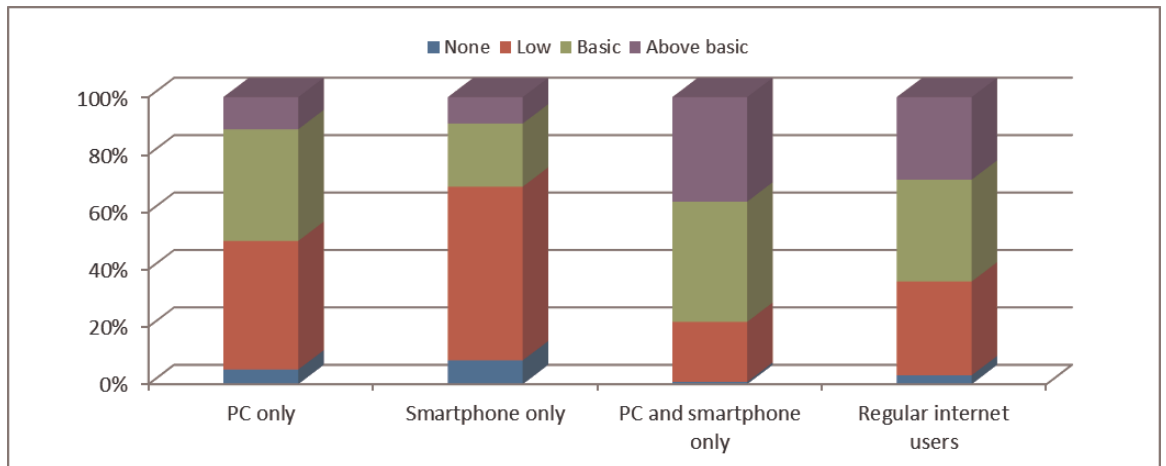


Figure 2 – Individuals aged 16-74 years regularly using the internet by overall level of digital skill (percentage values). Year 2016



Looking at the relation between the different devices and digital skills highlights that advanced skills are more common among individuals combining pc and smartphone (Fig. 2). The result holds for all the four digital competence domains. However, differences among generations persist also for multiple device alliterates: the higher competences are found among young people while gains are lower for older age groups. Internet users reporting the exclusive use of smartphone show the greatest disadvantage in terms of digital skills: 8.3% highlight no digital skills and 60.5% had low digital skills (with corresponding values of 0.8% and 20.9%, respectively, for those using pc and smartphone). Finally, individuals accessing internet only through their pc can be distinguished in two main groups: those with low and those with basic digital skills (44.8% and 38.9%, respectively).

References

1. Compaine, B. 2001. *The Digital Divide: Facing a Crisis or Creating a Myth?* MIT Press: Cambridge (Mass)
2. Di Maggio, P.J., Hargattai, E (2002). From the 'digital divide' to 'digital inequality' studying internet use as penetration increases. *Princeton: Center for Arts and Cultural Policy Studies, Woodrow Wilson School, Princeton University*, 4(1), 4-2.
3. Hargattai, E. (2002). Second Level Digital Divide: Differences in People's Online Skills, *First Monday* (7) 4
4. European Commission (2016). *The Digital Competence Framework 2.0* <https://ec.europa.eu/jrc/en/digcomp/digital-competence-framework>.
5. Istat (2016). Cittadini, imprese e nuove tecnologie. *Statistiche Report*, 21 dicembre



Section 4. Posters

Modelling Hedonic Price using semiparametric M-quantile regression

Regressione m-quantilica semiparametrica per la modellizzazione dei prezzi edonici

¹Riccardo Borgoni, Antonella Carcagni, Alessandra Michelangeli, Nicola Salvati

Abstract House prices depend on building, environmental, amenities and accessibility characteristics as well as urban and architectonic factors. In this paper, a semiparametric random effect M-quantile model is proposed to model the relationship between the house prices and its predictors for various M-quantiles of the distribution. In particular, the model includes a flexible component in the linear predictor along with a (urban) random component. The semiparametric term, which is expected to grasp the nonlinear relationship between prices and a cultural amenities index, is specified via a univariate spline. We treat the coefficients of the knots of the univariate spline as a further random component in order to obtain smoother results.

Abstract *I prezzi del mercato mobiliare dipendono da una serie di caratteristiche, architettoniche, ambientali e di contesto degli edifici. In questo lavoro viene proposto un modello di regressione M-quantilica semiparametrico per analizzare le relazioni fra i prezzi e tali caratteristiche. In particolare, la componente semiparametrica, specificata tramite spline penalizzate, è stata introdotta per cogliere effetti potenzialmente non lineari dell'indicatore di contesto culturale.*

Key words: M-quantile, spline, hedonic approach.

¹ Riccardo Borgoni, Department of Economics, Management and Statistics (DEMS) Università degli Studi di Milano-Bicocca, E-mail address: riccardo.borgoni@unimib.it
Antonella Carcagni, Department of Economics, Management and Statistics (DEMS) Università degli Studi di Milano-Bicocca, E-mail address: antonella.carcagni@unimib.it
Alessandra Michelangeli, Department of Economics, Management and Statistics (DEMS), Università degli Studi di Milano-Bicocca, E-mail address: alessandra.michelangeli@unimib.it
Nicola Salvati, Department of Economics and Management (DEM), Università degli Studi di Pisa, E-mail address: nicola.salvati@unipi.it

Introduction

In this paper, a semiparametric M-quantile random effects regression model (SMQRE from now on) is used to model the hedonic price equation for housing market. The hedonic approach interprets the price of a house as the market evaluation, by a hedonic price function, of the particular package of characteristics embodied in it (Gravel et al. 2006). Up to date, a strand of hedonic literature applied to the housing market has used quantile regression (Koenker and Basset, 1978) to capture consumer heterogeneity in housing demand or to identify housing sub-markets. The latter are usually defined as “geographical areas based on either pre-existing geographic or political boundaries (Amédée-Manesme et al. 2017).

This paper is the first attempt to use the M-quantile approach to model the distribution of the housing price conditional to housing attributes. An advantage of the M-quantile model lies in the fact that it links the robustness of quantile and the efficiency of expectile regression. Moreover, M-quantile regression is a flexible approach, which makes it possible to choose between a variety of continuous influence functions.

Hereafter, we propose an M-quantile regression model including two random terms allowing to control for both sub-market heterogeneity and non-linearity that may exist in the relationship between variables.

The suggested methodology is illustrated using data for the city of Milan, which is the largest city in Italy after Rome. The total population was 1,345,135 inhabitants in 2011, within an area of about 183 square km. The data come from different sources and are combined into a single data set. Housing market data are from the Real Estate Observatory (Osservatorio del Mercato Immobiliare). They provide information about 4,000 individual housing transactions in Milan between 2004 and 2010. In addition to housing market values, the data set provides information and detailed description of housing-specific attributes of the sample units, including total floor area, floor level, the presence of an elevator, whether housing unit has independent heating and the presence of a garage.

We retrieved geocoded data about local-specific characteristics from the open data portal of the municipality of Milan² and the Regional Environmental Protection Agency (ARPA) of the Lombardy region. In particular, we consider the availability of education, air quality, cultural activities and related infrastructures.

² <https://geoportale.comune.milano.it/sit>

The Semiparametric M-quantile random effects model

M-quantile regression (Breckling and Chambers, 1998) is a ‘quantile-like’ generalization of regression based on influence functions (M-regression) able to grasp differential effect of a covariate at different levels of the conditional distribution of the response variable. The M -quantile of order q for the conditional density of y given the set of covariates X , $f(y|X)$, is defined as the solution $MQ_y(q|X, \psi)$ of the integral equation $\int \psi_q [y - MQ_y(q|X, \psi)] f(y|X) dy = 0$ where ψ_q denotes an asymmetric influence function. In this paper The Huber loss function (Huber 1981) is used to define the linear M-quantile regression model.

The Real Estate Observatory divides the municipality of Milan in 55 administrative areas. Each area is internally homogeneous in terms of socioeconomic and urban characteristics, making it likely that prices of houses located within an area move together. Hence, the model considered hereafter includes a random intercept to account for a potential clustering of the data. To evaluate the social value of cultural activities in the city, we include a measure, termed the Cultural Catalyst in the rest of the paper, which captures the joint effect of cultural activities on housing market values (Borgoni et al 2018). The effect of this variable is expected to be not linear. Hence, to adjust for this potential non-linearity, a flexible component is added to the linear predictor of the M-quantile random effect model. In particular, we use penalised splines rely on a set of univariate quadratic basis functions to handle non-linear structures in the data. We treat the coefficients of the knots of the basis as a further random component in order to obtain smoother results (Ruppert et al. 2003).

The model we propose, for M-quantile q is:

$$MQ_y(X, Z, Z_{sp}; \psi) = \mathbf{X}\beta_q + \mathbf{Z}u_q + \mathbf{Z}_{sp}\gamma_q$$

where, Y is the house price taken on the log scale, \mathbf{X} is a $n \times (p + 1)$ matrix of the explanatory variables whose first column is constantly equal to 1 to include the intercept in the model, β_q is the $(p + 1)$ -vector of M-quantile specific regression coefficients. u_q is a G -vector of 0-mean Gaussian random effect and γ is a $K \times 1$ vector of 0-mean Gaussian random effects associated with the spline matrix independent of u_q . \mathbf{Z} is a $n \times G$ incidence matrix coding neighborhood hierarchy, whereas \mathbf{Z}_{sp} is a $n \times K$ spline matrix with entry (i, j) given by $[(x_i - k_j)]_+^2$ where k_j is the j -th knot of the spline, x_i the i -th sampling value of x , K the overall number of knots and $[x]_+$ is equal to x if $x > 0$ and 0 otherwise.

A robust maximum likelihood approach (Richardson and Welsh, 1995) has been adopted to estimate the model using the two-stage algorithm proposed by Tzavidis et al. (2016).

All the covariates included in the \mathbf{X} design matrix of the model are dummy variables except for the Cultural Catalyst. The total floor area is coded on three levels (< 63 ; $(63,115]$; $(115,491]$) and entered in the model via two dummy variables, the category $(63,115]$ being used as baseline. The \mathbf{X} matrix also includes a binary variable equal to 1 if the housing unit has been sold in the pre-crisis sample period (2004-2008) and 0 otherwise.

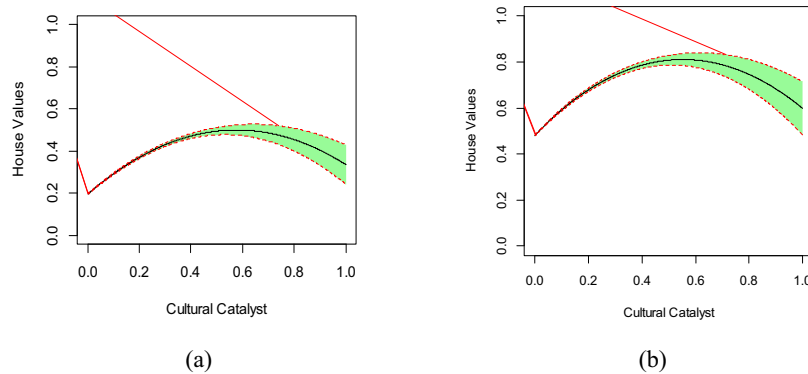
Results and conclusion

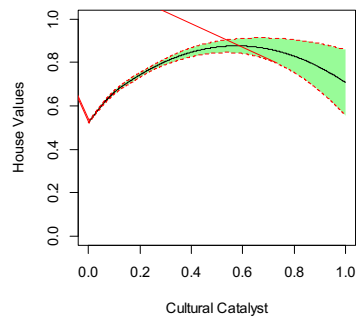
In this section, we briefly present the results obtained applying the SMQRE model to housing market prices of Milan. Many predictors representing the intrinsic attributes of the house are found to be statistically significant at all the investigated M-quantiles namely: total floor area, presence of the elevator, presence of a parking area, number of bathrooms, distance from university.

To account for the environmental quality of the place where houses are built, we add an air quality index and a dismissed area index to the regression. Both variables are also found statistically significant at all the considered quantiles.

In order to assess whether the expected non linear effect of Cultural Catalyst on housing prices is statistically significant, we construct a 95% confidence interval of the spline component following an approach akin to what Ruppert et al. (2003) suggested for usual semiparametric regression models. The result is reported in figure 1 for three different M-quantiles, namely $q = 0.25, 0.5, 0.75$.

Figure 1: Spline effect of the cultural catalyst at M-quantile (a) 0.25, (b) 0.50 and (c) 0.75





(c)

The confidence bands are located largely above zero for every value of the Cultural Catalyst suggesting a definitely significant effect of the cultural dimension on the house prices.

Finally, the floor where the dwelling is located positively contributes to determine the equilibrium-housing price only for high-priced apartments, whereas the period of the property crisis has a substantial negative impact only on lower M-quantiles.

Reference

1. Amédée-Manesme, C-O., Baroni, M., Barthélémy, F., des Rosiers F., (2017). Market heterogeneity and the determinants of Paris apartment prices: A quantile regression approach *Urban Studies*, 54, 14, 3260 – 3280.
2. Breckling, J., Chambers, R., (1998). M-quantiles. *Biometrika*, 75: 761–771.
3. Borgoni, R., Michelangeli, A., Pontarollo, N., (2018). The Value of Culture to Urban Housing Markets. *Regional Studies*, 52(12), 1672-1683.
4. Gravel, N., Michelangeli, A., Trannoy, A., (2006). Measuring the social value of local public goods: an empirical analysis within Paris metropolitan area, *Applied Economics*, 38, 1945-1961.
5. Richardson AM and Welsh AH. Robust restricted maximum likelihood in mixed linear models. *Biometrics* 1995; 51:1429–1439.
6. Ruppert, D., Wand, M., Carroll, R., (2003) *Semiparametric regression*. Cambridge University Press, Cambridge
7. Tzavidis, N., Salvati, N., Schmid, T., Flouri, E., Midouhas, E., (2016). Longitudinal analysis of the strengths and difficulties questionnaire scores of the millennium cohort study children in england using m-quantile random effects regression. *Journal of the Royal Statistical Society: Series A*, 179 (2), 427-452.
8. Huber, P., (1981). *Robust statistics*. Wiley, New York
9. Koenker, R., Bassett, G., (1978). Regression quantiles. *Econometrica* 46:33–50

Bayesian mixed latent factor model for multi-response marine litter data with multi-source auxiliary information

Modello bayesiano misto a fattori latenti per l'abbondanza di rifiuti marini con informazioni ausiliarie di diversa provenienza

Crescenza Calculli, Alessio Pollice, Marco V. Guglielmi and Porzia Maiorano

Abstract This work aims at analyzing the composition and spatio/temporal distribution of marine litter amounts found at the sea-floor in a region of the central Mediterranean. Inspired by common multi-species distribution modeling problems, we propose a suitable Bayesian multivariate approach to model litter data inferring possible environmental covariates while controlling for correlation between different litter categories and providing a method for residual ordination. A combined environmental information coming from multiple sources at different spatio/temporal scales is considered to investigate environmental drivers that might affect dynamics of marine litter assemblages at local scale.

Abstract L'obiettivo di questo lavoro è quello di analizzare la composizione e la distribuzione spazio/temporale dell'ammontare di rifiuti marini ritrovati sul fondale in una regione del Mediterraneo centrale. Traendo spunto dai modelli distributivi multi-specie, proponiamo un modello bayesiano multivariato per dati relativi ai rifiuti marini che tiene conto di possibili driver ambientali, considera la correlazione tra le differenti categorie di litter e permette l'ordinamento residuale. Al fine di individuare i fattori ambientali che influenzano le dinamiche di assemblaggio dei rifiuti marini a livello locale, è stata considerata una combinazione di informazioni ambientali provenienti da fonti differenti caratterizzate da diversa scala spazio/temporale.

Key words: Multi-species distribution model, Bayesian latent variable models, Spatio-temporal support, Marine litter data

Crescenza Calculli, Alessio Pollice

Department of Economics and Finance, University of Bari Aldo Moro, Largo Abbazia S. Scolastica, Bari, Italy, e-mail: crescenza.calculli@uniba.it, alessio.pollice@uniba.it

Marco Vito Guglielmi, Porzia Maiorano

Department of Biology, University of Bari Aldo Moro, LRU CoNISMa Bari, Via E. Orabona 4, Bari, Italy, e-mail: marcov.guglielmi@gmail.com, porzia.maiorano@uniba.it

1 Introduction

Recent progresses in monitoring networks and sensing systems provide remarkable opportunities to use large amounts of information to understand complex Ecological problems [1]. Data sets with multiple variables measured at different locations and/or at different time points are very common in Ecology and their analysis cannot ignore the association between the measurements at a given location and among measurements across locations. This work allows to explore possible preliminary solutions to model such data. In particular, we investigate the composition and the spatio/temporal distribution of waste amounts found at the sea-floor, combining environmental information from multiple sources with different spatio/temporal scales. Such environmental drivers can affect the dynamics of litter assemblages at the local scale. Marine litter has become a growing environmental concern that threatens life in seas, the status of coastal areas, and human health [2]. Marine waste results from anthropogenic activities on land and at sea (*e.g.* recreational activities, harbors or sewage-disposal sites, shipping by boat, fisheries, offshore industries), thus several environmental features are worth investigating. We propose to model the dependence of marine litter monitoring data on the effects of environmental covariates coming from various sources with different spatio/temporal support. As preliminary step, we pre-process environmental data to associate the target support (litter data) to various source supports (environmental data). A Bayesian mixed latent factor model, inspired by ecological multi-species distribution models [4], is thus proposed in order to investigate the distribution of marine litter in the area under study. This model-based approach merges univariate Generalized Additive Models (GAMs) with latent variables to account for the residual correlation across litter categories, *e.g.* due to environmental interactions or unaccounted covariates. This enables community-level inferences about environmental covariates while controlling for extraneous correlation between litter categories, as well as providing a method of residual ordination.

2 Marine litter data

Marine litter data are collected during experimental trawl surveys conducted from 2013 to 2016 in the North-Western Ionian Sea as part of the international project MEDITS (MEDiterranean International Trawl Surveys) activities. The study area represents the deepest sea in the Mediterranean basin characterized by a complex geomorphology and the presence of important fisheries and main harbors. In the last years, the sea bottoms of this area have been exposed to a strong anthropogenic impact due to the increasing tourist activity along the Ionian coasts. The same 70 geo-referenced depth-stratified hauls are sampled between 10 and 800 m in depth every year, summing to 280 hauls in 4 years. Wastes caught during the trawl surveys are classified in 8 categories: plastic, rubber, metal, glass/ceramic, cloth/natural fibres, processed wood, paper/cardboard, other/unspecified. The number of collected

items for each litter category was scaled to the swept surface unit (1km^2), thus obtaining density indices (N/km^2) for each litter category and survey at every haul location.

3 Environmental features

Some possible environmental drivers that might affect the spatio-temporal distribution of the multi-category litter density over the study region, are considered and discussed below.

- **Marine current.** Marine current data are retrieved from the CMEMS, Copernicus Marine Environment Monitoring Service (<http://marine.copernicus.eu/services-portfolio/access-to-products/>). For the Mediterranean area, the hydrodynamic-wave model supplied by the Nucleous for European Modelling of the Ocean (NEMO-OPA v3.4-3.6), combine available satellite and *in situ* data in order to produce geo-referenced data on sea water temperature, salinity and horizontal current. For the case study, we retrieved daily values of the eastern and the northward sea water velocity U and V at 1m depth over a horizontal grid with $6\text{-}7\text{km}$ resolution. The effect of marine currents is summarized averaging U and V for 8 days previous to sampling (suitable time lag to describe the effect of superficial marine currents) at the grid-point nearest to each haul.
- **Fishing activity.** Data related to fishing activities are retrieved from the Global Fishing Watch organization database (<https://globalfishingwatch.org/>). The service provides public access to vessel tracking data (identity, type, location, speed, direction and other information) broadcast using the Automatic Identification System (AIS). The daily fishing effort database allows to extract data for the North-Eastern Ionian Sea (GSA 19) with 0.01 degrees resolution for 3 vessel types: Drifted Longlines, Fixed Gears and Trawlers. For each grid cell and each vessel type, we consider the vessel annual transit and fishing time in hours. In order to align litter and fishing effort data, a radius of 5 km from each litter haul is considered. Within this radius, we obtain the annual average transit (MVH) and fishing time (MFH) per vessel considering the three ship types.

4 Mixed latent factor model for multivariate response

To model multivariate marine litter density, a mixed latent factor model for multivariate responses is proposed. As a preliminary analysis, we consider densities of the *plastic* category and densities of the aggregation of *all other categories*, leading to a bivariate response model. To account for response and covariates scale differences, the density of the two litter categories has been pre-scaled and assumed to be normally distributed $y_{ij} \sim N(\mu_{ij}, \tau)$. Thus the density μ_{ij} of the j -th litter category at the i -th haul, might be specified as follows

$$g(\mu_{ij}) = f_j(lon_i, lat_i) + \mathbf{X}_i \beta_j + \mathbf{b}_i + \mathbf{z}_i' \theta_j \quad i = 1, \dots, 280; \quad j = 1, 2 \quad (1)$$

where $g(\cdot) = \cdot$, $f_j(\cdot) = \sum_{h=1}^H d_h g_h(lon_i, lat_i)$ is a smooth function of haul coordinates where $g_h(\cdot)$ are spline-like basis functions at $H = 4$ knots controlled by the quadratic penalty $\sum_{h=1}^K \lambda_h \mathbf{d}' \mathbf{S}_h \mathbf{d}$ where \mathbf{S}_h 's are known coefficient matrices and λ_h are smoothing parameters to be estimated; $\beta_j = (\beta_{0j}, \dots, \beta_{Kj})$ is a vector of litter type-specific intercepts and $K = 4$ regression coefficients, \mathbf{b}_i are random annual effects, $\mathbf{z}_i = (z_{i1}, z_{i2})'$ is a vector of 2 latent variables accounting for response correlation and $\theta_j = (\theta_{j1}, \theta_{j2})$ is the vector of litter type-specific loadings. In this case, latent variables might be considered as missing informative predictors for the multivariate response inducing residual correlation between litter categories.

Independent weakly informative $N \stackrel{\text{iid}}{\sim} (0, 10)$ priors are assumed for the type-specific intercepts and regression coefficients, the random annual effects and the latent variables and loadings. Uniform priors $U \stackrel{\text{iid}}{\sim} (0, 30)$ are adopted for all dispersion parameters. For the smooth function, penalty is induced by improper multivariate Gaussian priors, $\mathbf{d} \sim N_H(\mathbf{0}, \mathbf{T})$ with $\mathbf{T} \propto \sum_h \lambda_h \mathbf{S}_h$ and standard gamma priors for smoothing parameters $\lambda_h \sim \Gamma(.05, .005)$ [5]. The multivariate approach is inspired by the R package `boral` [3] and Bayesian MCMC samples are obtained by the JAGS software and the `jagam` function of `mgcv` [6].

5 Main results

Model in Eq. (1) is fitted considering 4 environmental covariates (the current components U and V and the annual average transit and fishing time per vessel MVH and MFH), the random annual effects and 2 latent variables. The joint posterior distribution of model parameters were obtained using 12,000 iterations discarding the first 2,000. Chains were checked for convergence and reasonable mixing by graphical inspection of the trace plots and common convergence diagnostics (Gelman-Rubin diagnostics). Fig.1 shows estimated fixed effects distinguishing between litter categories: a relevant negative effect of the eastern current component (U) and a positive effect of the northern component (V) are obtained for the plastic density. This result suggests higher plastic densities with currents towards the North-western direction. Opposite signs are estimated for relevant effects of fishing effort covariates, implying higher plastic densities associated with an increasing vessel transit (MVH) but also lower densities with more intense fishing activity (MFH). Most likely, the result comes from the reduced number of plastic items caught by nets due to the heavy fishing activity in the area. No relevant effects are found for the density of other litter categories.

For both litter categories, the estimated spatial component in Fig. 2 shows the same areas with higher (Sicily and Calabria) and lower densities through the years.

Bayesian model for marine litter data

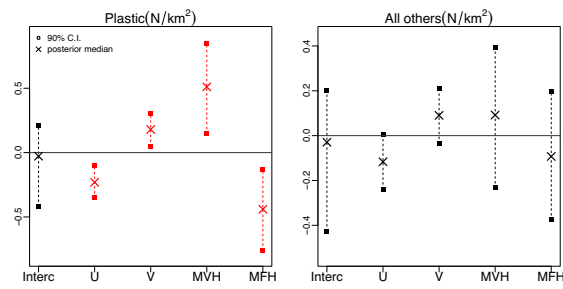


Fig. 1 Estimated fixed effects distinguishing between litter categories densities. Relevant effects in red.

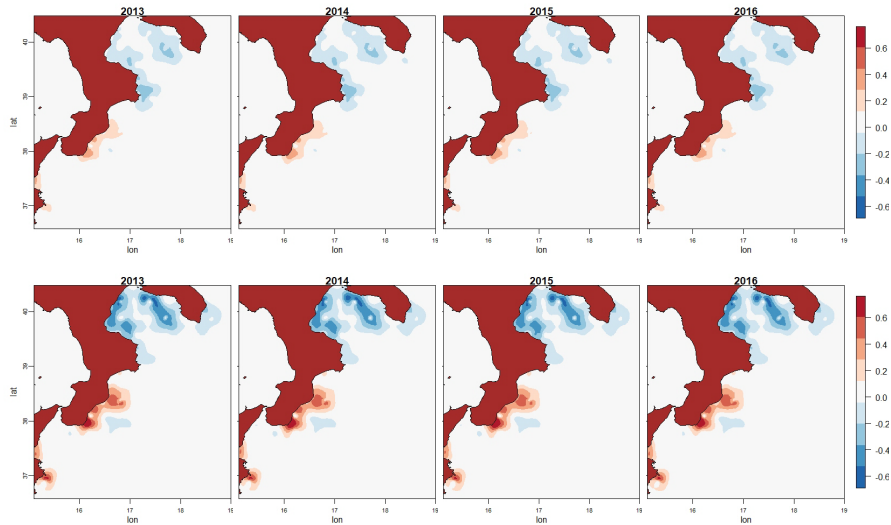


Fig. 2 Estimated residual spatial variation for plastic category (upper panel) and for all other categories (lower panel)

A more pronounced spatial effect for all other litter categories is possibly due to the absence of relevant covariate effects.

Biplots in Fig. 3 show the residual ordination obtained by estimated latent factors and allow to visualize the correlation between plastic and other litter categories: the two litter categories are strongly correlated between them and with the first axis ($\rho_{res} = 0.85$). The position of hauls in the biplots suggests that, over the years, a decreasing number of hauls are characterized by high levels of waste densities for both plastic and other litter.

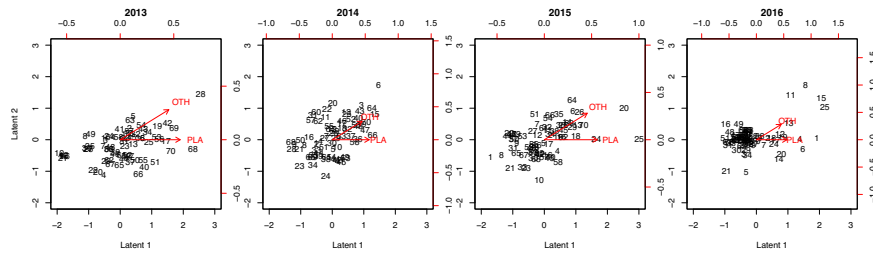


Fig. 3 Biplot of latent variable posterior medians. Hauls identify by numbers.

6 Conclusions and Further developments

This work represents a preliminary analysis of space/time structured multivariate marine litter data. Further developments include dealing with the change of support problem into the modeling framework, in order to combine data from multiple sources and different spatial/temporal sampling scales without the need of pre-process the data. The possibility to handle all individual litter categories and to include new environmental covariates not considered so far is also worth considering. Additional sources of auxiliary information include deeper currents and alternative definitions of the fishing effort.

Acknowledgements C. Calculi and A. Pollice were supported by the PRIN2015 project EphaStat - *Environmental processes and human activities: capturing their interactions via Statistical methods*, funded by the Italian Ministry of University and Research (MIUR). The MEDITS surveys have been carried out within the Data Collection Framework. For fishing activities data availability, we thank W. Zupa, COISPA Tecnologia & Ricerca, Bari, Italy.

References

1. Finley, A. O., Banerjee, S., Cook, B. D.: Bayesian hierarchical models for spatially misaligned data in R. *Methods in Ecology and Evolution* **5**, 514–523 (2014)
2. Gall, S.C., Thompson, R.C.: The impact of debris on marine life. *Marine Pollution Bulletin* **92**, 170–179 (2015)
3. Hui, F. K. C.: boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R. *Methods in Ecology and Evolution* **7**, 744–750 (2016)
4. Warton, D. I., Blanchet, F. G., O’Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., Hui, F. K. C.: So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology and Evolution* **30**, 766–779 (2015)
5. Wood, S.N.: Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC (2017)
6. Wood, S.: Just Another Gibbs Additive Modeler: Interfacing JAGS and mgcv. *Journal of Statistical Software* **75**, 1–15 (2016)

Official statistics to support the projects of A Scuola di OpenCoesione

L'esperienza di monitoraggio civico in Lombardia

nell'anno scolastico 2018-19

del Vicario G. and Di Gennaro L. and Ferrazza D. and Spinella V. and Viviano L.

Abstract In this work we want to describe the experience in support of Lombard schools involved in A Scuola di OpenCoesione (Asoc) through the use of official statistics data. For the first time in the 2018-19 school year, the Italian National Institute of Statistics (Istat) became part of the Asoc team. The Istat Territorial Office for Lombardy together with Europe Direct (ED, Asoc's historical partner) designed the training for the eight Lombard schools with classroom interventions and remote collaboration for the identification of data related to selected cohesion policy projects. The synergy allowed to work closely with the ED team and with teachers and students, promoting the culture of official statistics.

Abstract *In questo lavoro si vuole descrivere l'esperienza a supporto delle scuole lombarde coinvolte in A scuola di OpenCoesione (Asoc) attraverso il ricorso ai dati di statistica ufficiale. Per la prima volta nell'anno scolastico 2018-19 l'Istat è entrata a far parte del team di Asoc. L'Ufficio Territoriale Istat per la Lombardia insieme ad Europe Direct (ED, partner storico di Asoc) ha progettato la formazione alle otto scuole lombarde con interventi in aula e collaborazione a distanza per l'individuazione di dati legati ai progetti di politiche di coesione scelti. La sinergia ha permesso di lavorare a stretto contatto con il team ED e con i docenti e gli studenti promuovendo la cultura della statistica ufficiale.*

Key words: OpenCoesione, culture of official statistics, civic monitoring

¹

del Vicario Giusy, Istat- Istituto Nazionale di Statistica; email: delvicar@istat.it

Di Gennaro Luigi, Istat- Istituto Nazionale di Statistica; email: ludigenn@istat.it

Ferrazza Daniela, Istat- Istituto Nazionale di Statistica; email: ferrazza@istat.it

Spinella Valentina, Istat- Istituto Nazionale di Statistica; email: vspinella@istat.it

Viviano Lorena, Istat- Istituto Nazionale di Statistica; email: lovivian@istat.it

1 Cohesion policies

What cohesion policies are.

Cohesion policy is a multidimensional concept, whose objective is to promote sustainable territorial development among the Member States of the European Union.

Economic and social cohesion, as defined by the Single European Act of 1986, aims to "reduce the gap between the different regions and the backwardness of the least favoured regions". The most recent EU treaty, the Lisbon Treaty, adds a third dimension : "economic, social and territorial cohesion".

Cohesion policies refers to the policy framework which brings together thousands of projects across Europe receiving funding from the European Regional Development Fund (ERDF), the European Social Fund (ESF) and the Cohesion Fund (targeting EU Member States with a GDP below 90 % of the EU-27 average, not including Croatia).

These policies intervene in the territories to meet the specific needs of different places, in terms of infrastructure or services, but also human and social capital, in order to match their development opportunities [1].

1.1 The OpenCoesione portal

Web portal description.

OpenCoesione is the open government initiative on cohesion policies in Italy, coordinated by the Department for Cohesion Policies of the Presidency of the Council of Ministers. It allows citizens, administrators, technicians and entrepreneurs of innovation, researchers and journalists to have access to data and information, in order to assess the effectiveness and consistency of the use of resources of cohesion policies.

On OpenCoesione portal, it is possible to find out which projects are being financed, follow their progress and stimulate the planning and implementation processes through participation and reuse initiatives.

Data and information on interventions funded by cohesion policies in Italy can be accessed on the portal, they can be downloaded in open format, fed by additional resources from both European Funds, which is associated with national co-financing, and from National Funds. Open data on more than 1.183.365 projects has been posted so far with details on the amount of funding, policy objectives, locations, involved subjects and completion times. Data are updated every two months and can be freely re-used and explored interactively on the portal using maps, filters and detailed data sheets on projects and recipients [2].

1.2 Open cohesion School

Official statistics to support the projects of A Scuola di OpenCoesione
Open cohesion in Italian High Schools.

Open cohesion School¹ is the innovative project of experimental education started in the school year 2013-2014 aimed at high schools, which promotes principles of active and conscious citizenship by carrying out research and civic monitoring of public funding through the use of innovative information technologies, communication and data journalism, the development of digital skills and the use of data in open format (Open Data). It intends to help students to know and communicate how public communitarian policies intervene in places where they live, with the help of computer techniques, statistics and journalism.

In 2016-2017 edition 196 schools joined the project, with about 5,000 students and about 350 teachers.

Schools willing to join Asoc project identify a project financed according to different criteria: type, theme and territories. Two programming cycles are available: 2007-2013 and 2014-2020. The classes involved follow a training course with precise objectives and deadlines [3].

The educational path is divided into 5 lessons, a civic monitoring visit, active participation in the international event Open Data Day, a final public event. The didactics is organized according to a mixed model that provides methods of use of online content (such as MOOC) and project-based group work through the use of technologies, online sharing tools, blogs, social networks.

1.3 Istat and Asoc

National Statistics Institute's role in the project.

In the 2018-19 Istat joins the network of subjects supporting schools in the territories. Network are composed by the Europe Direct - EDIC Information Centres and the European Documentation Centres (EDCs), but also voluntary organizations and associations.

For the first time, therefore, Istat is one of the institutional partners that collaborate with the Asoc national team in order to strengthen the civic monitoring activities of the participating schools, providing Istat experts on specific issues addressed during the lessons provided by the project. The three years agreement at national level is then concretely implemented through the staff present in the regional Istat offices.²

The aim is to promote dissemination of statistical culture, to provide an adequate in-depth study of research methodologies and the construction of synthetic indicators and to enrich the training of students during their educational journey.

This activity is part of the promotion of statistical culture with the central coordination of the Central Directorate for the Development of Information and Statistical Culture (DCSI).

¹ Asoc is the Italian acronym for A Scuola di OpenCoesione.

² The central coordination of the project is managed by the Central Directorate for the Development of Information and Statistical Culture (DCSI) in ISTAT as project of promotion of statistical culture.

2 The experience of the Istat Lombardy Regional Office

The contribution of the Istat regional office

Within the Istat Lombardy Regional Office, a team of 5 members has been set up, including technical collaborators and researchers. Each of them put his/her skills to work and contributed to the reflection on general approach to be offered to schools.

The overall design of the Istat contribution to the Asoc project in Lombardy for the 2018-2019 school year is the result of a successful synergy with the Europe Direct Lombardy Region Office, which is the European Commission's information centre at regional level. The contents of the Istat interventions were shared and co-designed with the 'colleagues' of Europe Direct, in order to tailor them as much as possible to the needs of the classrooms, while guaranteeing the homogeneity of the general themes.

The Istat contribution is part of an organizational structure based on the experience of past years, bringing the added value of the official statistics point of view. The Asoc team at Istat Lombardy is composed by people with experience in the field of promoting statistical culture, used to interfacing with the public and in particular with the younger age groups with whom it is necessary to use an appropriate language and an engaging strategy. The challenge in the Asoc project was to 'test' the relevance of the official databases on the themes of the projects chosen by the schools, and therefore on important themes for the common welfare.

Istat goal in this project is to provide teachers and students involved in Asoc with the knowledge of official statistics, the appropriate use of statistical terminology, some insights on data acquisition through the construction of the statistical questionnaire, the reading of data tables and navigation of the main Istat databases.

Basically, Istat Lombardy offered an additional frontal lesson on basic statistical concepts and online resource (official statistics online dataset) tailored to projects objectives. We tried to explore official statistics dataset giving schools concrete examples of statistical indicators and variables related to their projects work.

An intervention scheme has therefore been defined, such as the one shown below, which has been repeated X times to X classes.

Table 1: Topic and training objectives identified for the lessons

<i>Topic</i>	<i>Training objectives</i>
Official Statistics: characteristics, actors and instruments (<i>high-level overview with references for more details</i>)	<i>Understand</i> the difference between official statistics, tout court statistics and NON-statistics; <i>Gain awareness</i> of the National Statistical System
Questionnaire: Structure-Dos and Don'ts- With/without Interviewer -Privacy (<i>high-level overview with references for more details</i>)	<i>Have an overview</i> useful for building a data collection tool
Basic statistical vocabulary:	<i>Acquire</i> a basic vocabulary to manage

Official statistics to support the projects of A Scuola di OpenCoesione

Number, Variable, Data, Metadata, Indicator, Index, Table elements	information rendered through the statistical data
Main graphs and their characteristics	<i>Acquire</i> basic elements to know how to construct the representation of the data (table and graph)
Istat and European dataset navigation	<i>Acquire</i> basic information on the various data sets made available by Istat; <i>Gain direct experience</i> of the navigation experience based on examples relevant to the project chosen by the school; <i>Get awareness about</i> comparability of data at an international level

The initial objective of the framework provided is to support the classes involved in the civic monitoring project for the drafting of an article in Data Journalism, with the identification of the most relevant official statistical data to be integrated with other sources, not necessarily referring to National statistical system (which Italian acronym is Sistan)

The awareness of the data and its correct interpretation, in general, if properly acquired, can represent a cognitive basis to be used for future studies, especially to bring out the critical spirit that every citizen should have against the data disseminated by a variety of channels daily.

The materials of the interventions have been prepared with the double objective of maintaining an agile and sustainable classroom situation (slides with little text and use of evocative images) and supporting texts to be delivered to teachers as a resource to be consulted ex post.

Table 2: Classes of eight schools in Lombardy participating in the school year 2018-19

Project	School	Keyword
Interventions for bike mobility to complete the existing network	Istituto Marie Curie – Cernusco sul naviglio (MI)	Sustainable viability; Lifestyle improvement; Cycle paths
Regio- implementation of an operational system for management of the procedures for administration of goods confiscated from organized crime	Istituto Benini – Melegnano (MI)	Confiscated goods; Redevelopment
The city in the city: a meeting place for people at Como	Liceo Gallio - Como	Families; Youth distress; Fragility; Solidarity
Total efficiency 4.0	Istituto Feltrinelli- Milano	R&D; Industry 4.0; Digitalization

Ferrazza D, Del Vicario G., Di Gennaro L, Spinella V., Viviano L.		
Digitalization of the high frequency network +valliTV	ITE AFM Federico Meneghini – Edolo (BS)	Digitalization; Competitiveness Telecommunications
Along the ridges: pathways between nature and culture for integrates development of the Camonica Valley	ITE Turismo Federico Meneghini – Edolo (BS)	Tourism
Lamination work on the river Lambro in the municipalities of Inverigo, Nibionno and Veduggio con Colzano	Liceo delle scienze umane Parini-Seregno (MB)	Environment Energy;Flood
Minicipality of Sernio	ITE SIA Pinchetti-Tirano (SO)	Tourism

3 Future developments

Evaluation and next step.

The project of civic monitoring of schools will end in May 2019. Winners will be offered a trip to the main European institutions in Brussels.

At the end of the projects, Istat Lombardia will start a process of evaluation of its activities, involving local stakeholders and schools.

The results will be used to refine the activities for the coming years, during which we hope to strengthen synergies with other stakeholders and to continuously improve in supporting the younger generations to understand the importance of official statistics to help to take decisions and the importance to be aware citizens.

References

1. Cohesion policies: https://ec.europa.eu/regional_policy/it/faq/#1
2. OpenCoesione Portal: <https://opencoesione.gov.it/it/>
3. Open cohesion School or A Scuola di OpenCoesione: <http://www.ascuoladiopencoesione.it/>

Spatial Logistic Regression for Events Lying on a Network: Car Crashes in Milan

Regressione logistica per eventi su network: gli incidenti automobilistici nel comune di Milano

Andrea Gilardi, Riccardo Borgoni and Diego Zappa

Abstract In this paper we propose a methodology to estimate the probability that a car accident occurs in urban roads. Our approach is based on logistic regression and takes into account the particular nature of the data which conforms to a spatial point pattern on a network. Using the open data on street networks provided within the OpenStreetMap project, we estimate the probability of car accidents for every street in the municipality of Milan.

Abstract *In questo lavoro viene presentato un approccio basato sulla regressione logistica per stimare la probabilità che avvenga un incidente automobilistico sui tratti di strada urbani. La metodologia proposta tiene conto della natura dei dati disponibili, assimilabili ad eventi casuali su un supporto spaziale di network. Utilizzando gli open data disponibili dal progetto OpenStreetMap, la stima è ottenuta per la totalità delle strade presenti sul territorio comunale di Milano.*

Key words: urban geography, car accidents, open data

1 Introduction

A precise definition of Smart City is extremely difficult. Nevertheless, one of the key point that several authors agree upon is that a smart city should use modern technologies to improve urban traffic and street safety [8]. During the last years the European commission launched a new campaign to improve the transportation

Andrea Gilardi
University of Milano - Bicocca, e-mail: a.gilardi5@campus.unimib.it

Riccardo Borgoni
University of Milano - Bicocca

Diego Zappa
Catholic University of Milan

system, promote sustainable mobility and avoid road congestion, accidents and air pollution [2].

In this paper we estimate the probability that at least one car accident occurs for every street in Milan considering a dataset reporting the spatial location of the accidents.

Hereafter, it is assumed that the occurrence of accidents in the city area conforms to a spatial point process. Since the likelihood of a planar Poisson point process can be approximated by the likelihood of logistic regression for the discretised process, pixel-based logistic regression is now commonplace to analyse the spatial dynamic of a point pattern in applied statistics and GIS literature. The connections between the two approaches are thoroughly discussed in [1]. In addition, resorting to a GLM model is particularly convenient when one is interested in the potential effect of a number of covariates.

Car accidents, however, are a classical example of a point pattern occurring on a linear network [4]. In these circumstances, the usual statistical techniques designed for point patterns occurring on the plane can not work since it is necessary to take into account the fact that the events are constrained to lie on the network.

In this paper, we propose a spatial logistic regression to model network binary data akin to the pixel-based logistic regression approach. However, given the nature of the process, we abandon the idea of dividing the space into pixels. Instead, we split the street network of Milan into smaller segments recording the presence of car crashes into each segments and fit a logistic regression model.

We underline the fact that, whenever it was possible, we chose to use open access data. They are becoming more and more popular in the road safety research methods (e.g. [6] and [3]) but they are still of very limited use in the Municipality of Milan. Nevertheless, we strongly believe in their importance for a digital, sustainable and safe development of a Smart City, as it was suggested by [7].

2 Data and methods

We consider accidents occurred between the 1st of January and 31st of December 2015, that require an ambulance intervention and for which the exact time and location, geocoded in UTM coordinates, were recoded. The sample includes 8601 events. Their annual temporal distribution is reported in Figure 1(a) and their hourly temporal distribution is reported in Figure 1(b). Data show some seasonality. Firstly, the number of interventions per day exhibits some peaks in winter (in particular during the Christmas holidays) and a drastic reduction in the last days of July and in the first days of August (at the beginning of summer holidays). Secondly, the hourly temporal distribution of car crashes is deeply different between working days and weekends.

The road network was built using data from OpenStreetMap [5], which is a project that aims to build a free and editable map of the world with an open-content license. The basic components of OpenStreetMap data are called *elements* and they

consist of: *nodes* (related to points on the earth surface), *ways* (which is a list of nodes and the most important structure for our model) and *relations* (describing the interactions between nodes and ways). Every physical object in the landscape is represented by these three elements and its attributes are stored using a `tag`, which is simply a pair of items which identify a category, a `key`, and the corresponding value¹, for instance `street=motorway` or `name="Viale Sarca"`. It is important to point out that almost all streets are internally stored as the union of a set of *segments* and the logistic regression model presented later on in the paper is based on this particular segmentation.

We downloaded data for all *trunk*, *primary*, *secondary*, *tertiary*, *unclassified*, *residential* and *service* highways for a total of 34085 segments with different length. Although there is no exact relationship between OpenStreetMaps tags and Italian Administrative road classification, the selected levels range from *Strada statale* to *Strada vicinale*. In order to apply a spatial logistic regression model, we projected all car crashes to the nearest point belonging to the linear network and the final result is reported in Figure 2.

For each segment of a street, a binary response variable Y is defined taking value 1 if at least one car accident occurred in the segment and 0 otherwise. Given the high granularity of the network, this procedure resulted in a slightly imbalanced response variable with 30149 *zeros* and 3936 *ones*. Notice that, in the rare cases (approximately 1 every 200 interventions) when a car accident got projected exactly to the common boundary point of two touching segments, then it was assigned to both of them.

The covariates included in the logistic regression are: the segment length, its OpenStreetMap classification, and the total number of crossed segments, the latter used as a proxy of urban traffic. A logarithm transformation is applied to the segment length, given the strong skewness of its distribution.

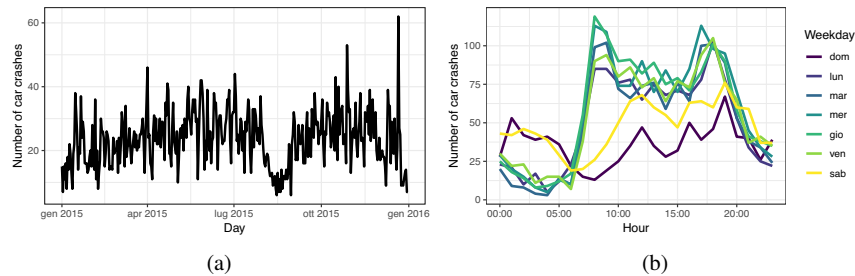


Fig. 1: Temporal distribution of ambulance interventions in case of car crash per day of the year (a) and per hour and day of the week (b) during 2015

¹ A complete list of all the keys and the corresponding values can be found at https://wiki.openstreetmap.org/wiki/Map_Features

Using the iteratively reweighted least squares (IWLS) algorithm implemented in the R function `glm`, we fitted the following logistic regression model

$$\text{logit}(\mathbb{P}(Y = 1|X_1, X_2, X_3)) = \beta_0 + \beta_1 \log(X_1) + \beta_2 X_2 + \beta_3 X_3, \quad (1)$$

where $\text{logit}(p) = \frac{p}{1-p}$, X_1 denotes the segment length, X_2 is a categorical variable representing its OpenStreetMap classification and X_3 is the number of touching segments.

3 Results and conclusions

The main results obtained applying the methodology presented above to the car accident data of Milan can be summarised in the following points:

1. the length of the segment is found to be the most important determinant to evaluate its risk with longer segments having an higher probability of car crashes;
2. *Secondary* and *Tertiary* segments (partially corresponding to *Strada Statale* and *Strada Provinciale*) are classified as the most dangerous roads regarding car crashes.
3. the number of touching segments does not influence the probability of car crashes, which probably means we are just using a bad proxy for urban traffic.

A graphical representation of the estimated probability of car crashes is reported in Figure 3. The segments far from the city centre exhibit higher probability of car

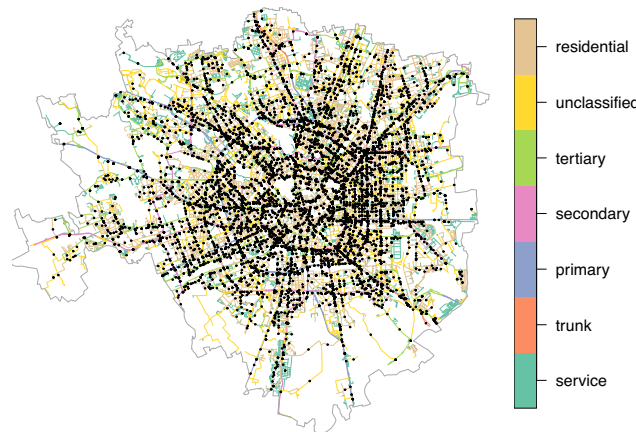


Fig. 2: Spatial representation of all car crashes in Milan during 2015 that required the intervention of an ambulance. Every street is colored according to its OpenStreetMap classification.

accident. The goal of this study is to propose a procedure to estimate a car accident risk index for every street in the Municipality of Milan taking into account the network nature of the spatial support of the data appropriately. Nevertheless this is just a preliminary result towards the development of a safety index and several enhancements are possible, in particular regarding the data and modelling perspective. Firstly it may be relevant to add demographic (e.g. population density), economic (e.g. number of vehicles per household) and traffic (e.g. total network volume and average travel speed) variables to the logistic regression model to improve its predictive performances. Secondly, alternative specifications for the link function could be considered. Finally, the classic independence assumption of GLM models can be somewhat questionable for this particular application. Hence, a reasonable extension of the model proposed in this paper is to take into account the potential spatial autocorrelation of adjacent segments.

Acknowledgements Map data copyrighted OpenStreetMap contributors and available from <https://www.openstreetmap.org>.

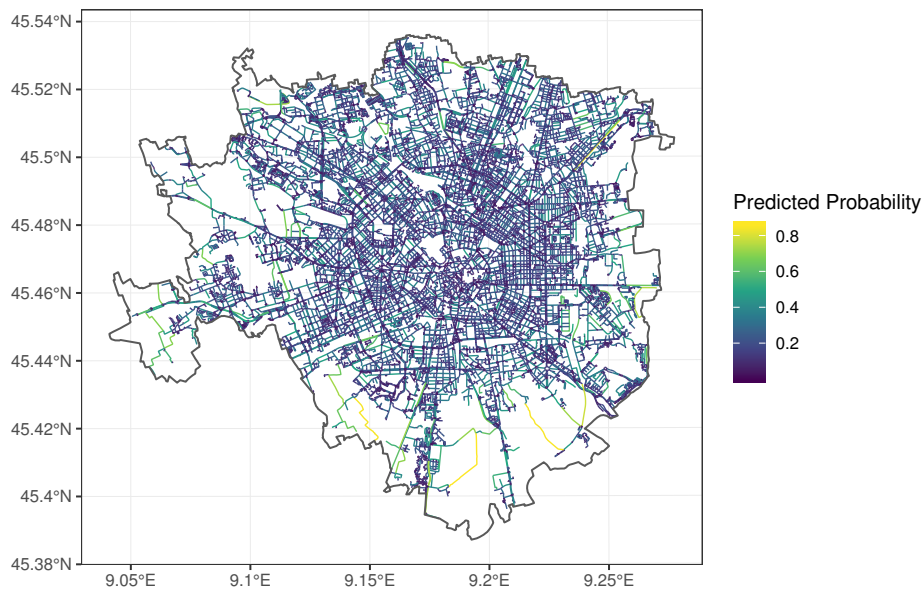


Fig. 3: Estimated probability of accident occurrence of the street segments displayed in Figure 2 based on a logistic regression model. Segments far from the city centre exhibits higher probability of car crashes.

References

1. Baddeley, A.; Berman, M.; Fisher, N.I.; Hardegen, A.; Milne, R.K.; Schuhmacher, D.; Shah, R.; Turner, R. Spatial logistic regression and change-of-support in Poisson point processes. *Electron. J. Statist.* 4 (2010).
2. European Commission: European Initiative on Smart Cities, 2010–2020. <http://setis.ec.europa.eu/set-plan-implementation/technology-roadmaps/european-initiative-smart-cities>
3. Lovelace et al., (2019). stats19: A package for working with open road crash data. *Journal of Open Source Software*, 4(33), 1181.
4. Okabe A, Sugihara K (2012) *Spatial Analysis along Networks: Statistical and Computational Methods*. Wiley, Chichester.
5. OpenStreetMap contributors: Planet dump retrieved from <https://planet.openstreetmap.org> (2017). <https://planet.openstreetmap.org>
6. Chinmoy, S.; Chris, W. & Sarika, K. (2018) Street morphology and severity of road casualties: A 5-year study of Greater London, *International Journal of Sustainable Transportation*, 12:7, 510-525.
7. Tuba, B.; Esteve, A.; Jonathan, W. (2013), A Smart City Initiative: the Case of Barcelona, *Journal of the Knowledge Economy*, 4, (2), 135-148.
8. Vito, A.; Berardi, U.; Dangelico, R. (2015) Smart Cities: Definitions, Dimensions, Performance, and Initiatives, *Journal of Urban Technology*, 22:1, 3-21.

Variable selection and classification by the GRID procedure

Selezione e classificazione delle variabili attraverso il metodo GRID

Francesco Giordano, Soumendra Nath Lahiri and Maria Lucia Parrella

Abstract We describe a method for variable selection and classification for a non-parametric regression in high dimensions where only a relatively small subset of variables are relevant and may have nonlinear effects on the response. The new method, called the GRID, is proposed and deeply investigated in a forthcoming paper. It is an extension of the RODEO method of [3] (which only makes variable selection). Among the novelties of our procedure, a graphical tool for identifying the low dimensional nonlinear structure of the regression function is shown. Given the length of this paper, we briefly describe the method and present the theoretical foundations and simulation performance of only the first stage of the procedure (*i.e.*, variable selection and linear/nonlinear classification).

Abstract *Descriviamo un metodo per la selezione e classificazione delle variabili di un modello di regressione nonparametrico in alta dimensione, dove solo un sottoinsieme relativamente piccolo di variabili è rilevante e può avere effetti non lineari sulla variabile risposta. Il metodo descritto, chiamato GRID, è proposto e analizzato in un paper in via di pubblicazione, e rappresenta un'estensione del metodo RODEO di [3] (che si limita alla selezione delle variabili). Tra le novità della nostra procedura vi è uno strumento grafico per rappresentare e identificare la struttura di bassa dimensione della funzione di regressione. Presentiamo qui una breve descrizione del metodo e i fondamenti teorici e una breve simulazione limitatamente al primo stadio della procedura (cioè selezione e classificazione delle covariate tra lineari e non lineari).*

Key words: Variable selection, nonparametric regression, high dimension.

F. Giordano
University of Salerno, Via Giovanni Paolo II, 132 e-mail: giordano@unisa.it

S.N. Lahiri
North Carolina State University, 5109 SAS Hall 2311 Stinson Dr, Raleigh, NC 27695-8203 e-mail: snlahiri@ncsu.edu

M.L. Parrella
University of Salerno, Via Giovanni Paolo II, 132 e-mail: mparrella@unisa.it

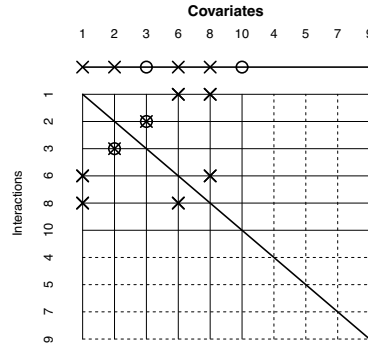


Fig. 1 GRID representation of model (2). The first row represents the selected linear and nonlinear variables (denoted by \times and \circ , respectively). The entries in the rest of the grid represent interactions among the selected variables (linear with linear and/or linear with nonlinear).

1 The GRID method

In this paper we describe the main ideas for a new method, called the *Gradient relevant identification of derivatives* (or the *GRID*) method, for simultaneous variable selection, classification of the relevant covariates between linear and nonlinear, and estimation of the low-dimensional structure of the regression function. This method is proposed and deeply investigated in a forthcoming paper by [2], and it is an extension of the RODEO method proposed by [3]. To briefly describe the methodology, consider the nonparametric regression model

$$Y_t = m(X_t) + \varepsilon_t, \quad t = 1, \dots, n, \quad (1)$$

where the X_t represents the \mathbb{R}^d -valued covariates and the errors ε_t are *iid* with zero mean and variance σ^2 . The errors ε_t are independent of X_t , and are assumed to be Gaussian, as in [3]. Here $m(X_t) = E(Y_t|X_t) : \mathbb{R}^d \rightarrow \mathbb{R}$ is the multivariate conditional mean function. We use the notation $X_t = (X_{t1}, \dots, X_{td})$ to refer to the covariates. We assume that the number of covariates $d \rightarrow \infty$ but only r of these covariates are relevant for model (1), where $r \ll d$ is bounded.

The acronym GRID has a two-fold meaning: first, it derives from *Gradient Relevant Identification of Derivatives*, meaning that the procedure is based on testing the significance of partial derivative estimators (derived by the Local Linear Estimation methodology); second, it refers to a graphical tool which can help in identifying the structure of model (1). Let $d = 10$ and let the true model be given by

$$Y_t = X_{t2} \log(X_{t3}) + X_{t1} X_{t6} X_{t8} + \exp(X_{t10}) + \varepsilon_t. \quad (2)$$

The graphical representation of the GRID methodology produces the plot in Fig. 1.

The first stage of the GRID procedure identifies (the indices of) the following sets of covariates (suitable to draw the top line of the GRID plot in Fig. 1).

$$C = \{3, 10\}, \quad A = \{1, 2, 6, 8\}, \quad U = \{4, 5, 7, 9\}.$$

The second stage of the GRID procedure derives (the indices of) the following sets of interactions

$$I^1 = \{1, 6, 8\}, \quad I^2 = \{2, 3\}, \quad I^3 = \{2, 3\}, \quad I^6 = \{1, 6, 8\}, \quad I^8 = \{1, 6, 8\}, \quad I^{10} = \{10\},$$

where I^j includes the interactions of variable j with other covariates. By default, each set I^j automatically includes the index j (self-interaction). Therefore, if the set I^j has the only component j , then X_{I^j} appears in the model as an isolated additive covariate, like $X_{I^{10}}$ in model (2). These index sets can be determined by successive scans of the columns of the GRID plot. Therefore, unlike existing methods, GRID identifies the interaction terms of any order among the relevant variables without having to pre-specify the highest order *a-priori*.

Next note that using the sets A , C and I^j from Figure 1, we can identify the actual low dimensional structure of the regression function $m(\cdot)$ as:

$$m(x_1, \dots, x_{10}) = \ell_1(x_1, x_6, x_8) + x_2 \cdot m_1(x_3) + m_2(x_{10})$$

for some nonlinear functions $m_1(\cdot)$ and $m_2(\cdot)$ and some (multi-)linear function $\ell_1(\cdot)$. To see this, it is sufficient to scan the GRID plot of Figure 1 by columns, reading each column from the top to the bottom. For example, looking at the column with label 10 ($j = 10$) we can note a circle at the top (in row 0), showing that $X_{I^{10}}$ is a nonlinear covariate, and no symbols below, indicating that x_{10} appears in $m(\cdot)$ as a nonlinear term by itself, with no interaction. Now consider the column $j = 2$, which shows a cross at the top row and one symbol in the column below it. This means that covariate X_{I^2} is linear and interacts with another covariate. The symbol in the column below is in positions 3, which corresponds to a nonlinear covariate (the particular symbols cross-circle shows the kind of interaction, which is linear with nonlinear). Therefore the interaction term involving variable X_{I^2} must be of the form $x_2 \cdot m_1(x_3)$. Column $j = 3$ of the plot reaffirms the interaction term involving the nonlinear variable X_{I^3} and the linear variable X_{I^2} . Similarly, the columns 1, 6 and 8 jointly yield the second (linear interaction) term of $m(\cdot)$ above. Note that the maximum order of the interaction terms corresponds to the max number of symbols appearing in the columns of the GRID plot (including position zero on the top). It is three in the case shown in Figure 1 for model (2).

2 Theoretical basis for the first stage of the GRID algorithm

Local linear estimation (LLE) is a nonparametric method for estimating the regression function $m(\cdot)$ in (1) (cf. [4]). To estimate $m(\cdot)$ at $x = (x_1, \dots, x_d)$, the LLE

performs a locally weighted least squares fit of a linear function. Let

$$\hat{\beta}(x; H) \equiv \arg \min_{\beta_0, \beta_1} \sum_{t=1}^n \{Y_t - \beta_0 - \beta_1^T (X_t - x)\}^2 K_H(X_t - x), \quad (1)$$

where the function $K_H(u) = |H|^{-1} K(H^{-1}u)$ gives the local weights with a d -variate product Kernel function $K(u) = \prod_{j=1}^d K_1(u_j)$. The bandwidth matrix H controls the bias and the variance of the resulting LLE of $m(x)$. For simplicity, we shall suppose that $H = \text{diag}(h_1, \dots, h_d)$ is a diagonal matrix with strictly positive entries. The estimator $\hat{\beta}(x)$ can be written in a closed form as:

$$\hat{\beta}(x; H) = (\Gamma^T W \Gamma)^{-1} \Gamma^T W Y, \quad (2)$$

where $Y = (Y_1, \dots, Y_n)^T$ and

$$\Gamma = \begin{pmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{pmatrix}, \quad W = \begin{pmatrix} K_H(X_1 - x) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & K_H(X_n - x) \end{pmatrix}.$$

Note from (1) that $\hat{\beta}(x; H)$ gives estimators of the function $m(x)$ and its gradient:

$$\hat{\beta}(x; H) = \begin{pmatrix} \hat{\beta}_0(x; H) \\ \hat{\beta}_1(x; H) \end{pmatrix} \equiv \begin{pmatrix} \hat{m}(x; H) \\ \hat{\mathbb{D}}(x; H) \end{pmatrix}. \quad (3)$$

For the theoretical derivations, we shall use the following assumptions:

- A1) The bandwidth H is a diagonal matrix with strictly positive diagonal entries: $H = \text{diag}(h_1, \dots, h_d)$, with $c_1 \leq h_j$ for $j = 1, \dots, d$ for some $c_1 \in (0, \infty)$.
- A2) The d -variate kernel function K is a product kernel, based on a nonnegative and symmetric univariate kernel density function $K_1 \in \mathcal{C}^1[-c_2, c_2]$ for some $c_2 > 0$ such that $0 < x_j - c_2 h_j < x_j + c_2 h_j < 1$ for all $j = 1, \dots, d$.
- A3) All the partial derivatives of the function $m(x)$ up to and including order five are bounded.
- A4) X_1 is uniformly distributed on the unit cube $(0, 1)^d$.

We now briefly comment on the assumptions. Assumption A1 requires the bandwidth matrix to be a diagonal matrix, which simplifies the development of the GRID method and is adequate for our purpose. The major difference between A1 and the typical assumption made on the bandwidth matrix H is that here the bandwidths h_j do *not* go to zero with the sample size (compare with [4]). As a consequence, the bandwidths do not behave like tuning parameters for the GRID procedure (see also [1] and [3]). The conditions on the d -variate kernel K and the design points in Assumptions A2, A3 and A4 are the same as in [3]. In particular, A2 implies that all the moments of the kernel K exist and that the odd-ordered moments of K and $(K)^2$ are zero.

Theorem 1. Under model (1) and assumptions A1-A4, the following result holds:

$$E \left\{ \frac{\partial \hat{m}(x; H)}{\partial h_j} \middle| \chi_n \right\} = \begin{cases} \theta_{0j} + o_p(1) & \text{if } j \in C \\ o_p(1) & \text{otherwise} \end{cases} \quad (4)$$

for $j = 1, \dots, d$, where $\chi_n = (X_1, \dots, X_n)$ and $\theta_{0j} \neq 0$. The exact expressions for θ_{0j} can be derived similarly as in [2].

Note that Theorem 1 can be used to identify only the nonlinear covariates, but not the linear covariates (those in the sets A or U). To overcome this, [3] suggest identifying first the linear variables through LASSO or to change the degree of the local polynomial estimator to zero (i.e. to use the Nadaraya-Watson estimator). Both these solutions make use of extraneous methods and require further attention to the choice of critical tuning parameters (the choice of the regularization parameter in the case of LASSO and the choice of the bandwidth matrix for the Nadaraya-Watson estimator) which are not well studied in the present framework. In comparison, here we propose a simple solution to the problem that allows the user to apply the same algorithm, but to a modified regression problem. Specifically, we consider an auxiliary regression where all those covariates that have not been selected in the first pass, are to be transformed, so that the *linear covariates* of the original model become *nonlinear* in the auxiliary model.

Define the transformation $z = \phi(x)$ and its inverse $x = \phi^{-1}(z)$ as follows:

$$z = \phi(x) = (x_C, x_{C^c}^{1/2}) = (x_C, x_A^{1/2}, x_U^{1/2}), \quad x = \phi^{-1}(z) = (x_C, z_{C^c}^2) = (x_C, z_A^2, z_U^2), \quad (5)$$

where C^c denotes the complement of C , and let $Z_t = \phi(X_t)$ be the transformed random variables. Next, using again the LLE technique as in (1)-(3), consider the following auxiliary regression

$$Y_t = m(\phi^{-1}(Z_t)) + \varepsilon_t \equiv g(Z_t) + \varepsilon_t, \quad t = 1, \dots, n.$$

Note that we use the same index partition as in the first regression. By (5), the covariates in A have a nonlinear effect in the auxiliary regression model $g(z)$. In fact, $z_j = \phi(x_j) = x_j^{1/2} \Rightarrow x_j = \phi^{-1}(z_j) = z_j^2$ for all $j \in A \cup U$, so the partial derivatives are

$$\frac{\partial g(z)}{\partial z_j} = \begin{cases} 2c_j z_j \neq 0 & \text{for } j \in A \\ 0 & \text{for } j \in U \end{cases},$$

where $c_j = \partial m(x)/\partial x_j$ is constant with respect to x_j , for all $j \in A$. Therefore, the linear covariates in A behave nonlinearly in the auxiliary regression, while the irrelevant covariates still remain so. Now we can state the following theorem to make linear variable selection from the covariates in the set C^c identified by Theorem 1.

Theorem 2. Suppose that Assumptions A1-A4. Then, with the transformed random variables $Z_t = \{\phi(X_t)\}$ of (5), the following result holds:

$$E \left\{ \frac{\partial \hat{g}(z; H)}{\partial h_j} \middle| \chi_n \right\} = \begin{cases} \theta_{0j}^g + o_p(1) & \text{if } j \in A \\ o_p(1) & \text{if } j \in U \end{cases} \quad (6)$$

Model 1		$d = 20$		$d = n/2$		$d = 2n$		Model 2		$d = n/2$		$d = 2n$	
Covariate	n	R	C	R	C	R	C	Covariate	n	R	C	R	C
X_2	300	0.805	0.805	0.530	0.530	0.300	0.303	X_1	300	1.000	*	0.995	*
	500	0.970	0.970	0.835	0.835	0.520	0.520		500	1.000	*	1.000	*
	1000	1.000	1.000	0.990	0.990	0.715	0.715		1000	1.000	*	1.000	*
X_3	300	0.920	*	0.710	*	0.450	*	X_2	300	0.855	*	0.560	*
	500	0.990	*	0.950	*	0.690	*		500	0.985	*	0.795	*
	1000	1.000	*	1.000	*	0.795	*		1000	0.995	*	0.905	*
X_4	300	0.915	*	0.790	*	0.430	*	X_7	300	1.000	0.665	0.970	0.645
	500	0.990	*	0.950	*	0.610	*		500	1.000	0.880	0.990	0.875
	1000	1.000	*	1.000	0.035	0.815	*		1000	1.000	0.995	1.000	0.995
X_5	300	1.000	*	1.000	*	0.990	*						
	500	1.000	*	1.000	*	1.000	*						
	1000	1.000	*	1.000	0.030	0.990	*						

Table 1 Simulation results for the two models in (7), with different dimensions d and sample sizes n . The values show the proportion of times that a given covariate X_{it} is classified as a relevant covariate (R) or as a nonlinear covariate (C). By difference of the two sets we can derive the set A . The symbol (*) denotes a value ≤ 0.025 . For all the irrelevant variables we observed a value (*).

for $j \in C^c$, where $\chi_n = (X_1, \dots, X_n)$ and $\theta_{0j}^g \neq 0$. For θ_{0j}^g see [2].

These two theoretical results allow us construct an appropriate threshold to select the variables, as in [3].

3 Some simulation results

The Monte Carlo simulation is based on 200 iterations. The covariates are uniformly distributed. We consider the two models $Y_t = m(X_t) + \varepsilon_t$ with

$$(1): m(x) = \sin(10x_2) + x_3x_4 + x_5 \quad (2): m(x) = x_1x_2 + x_1x_7^3 \quad (7)$$

and $\varepsilon_t \sim N(0, 1)$ for all t . The additive components of the models are standardized so that they all have variance equal to one, to make them comparable each other. The Kernel function is $K_1(u) = 1/C_1 (5 - u^2) \mathbb{I}_{\{|u| \leq \sqrt{5}\}}$, as in [3] and [2], where C_1 is a scale factor to make the integral equal one. The simulation results are shown in Table 1.

References

1. Bertin K., Lecue G.: Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics* **2**, 1224–1241 (2008)
2. Giordano F., Lahiri S.N., Parrella M.L.: GRID: A variable selection and structure discovery method for high dimensional nonparametric regression. To appear on *The Annals of Statistics*.
3. Lafferty J., Wasserman L.: RODEO: sparse, greedy nonparametric regression. *The Annals of Statistics* **36**, 28–63 (2008)
4. Ruppert D., Wand P.: Multivariate locally weighted least squares regression. *The Annals of Statistics* **22**, 1346–1370 (1994)

Joint VaR and ES forecasting in a multiple quantile regression framework

Stima congiunta del VaR e dell'ES attraverso la regressione quantilica multipla

Merlo Luca, Petrella Lea and Raponi Valentina

Abstract An accurate assessment of tail dependencies of financial returns is key for risk management and portfolio allocation. In this paper we consider a multiple linear quantile regression setting for joint prediction of tail risk measures, namely Value at Risk (VaR) and Expected Shortfall (ES), using a generalization of the Multivariate Asymmetric Laplace distribution. The proposed method permits simultaneous modelling of multiple conditional quantiles of a multivariate response variable and allows to study the dependence structure among financial assets at different quantile levels. Subsequently, we introduce a method for portfolio construction where we show that the portfolio returns follow a univariate asymmetric Laplace density. An empirical application to weekly returns of three major stock market indices, will be considered in the future to illustrate the practical applicability and relevance of joint estimation of VaR and ES in a multivariate framework.

Abstract Una stima accurata della dipendenza nelle code dei rendimenti finanziari fondamentale per la gestione del rischio. In questo paper proponiamo un modello di regressione quantilica multipla per la stima congiunta di misure di rischio, quali il *Value at Risk* (VaR) e l'*Expected Shortfall* (ES) sfruttando una generalizzazio-

Merlo Luca

Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, e-mail: luca.merlo@uniroma1.it

Petrella Lea

MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano 9, e-mail: lea.petrella@uniroma1.it

Raponi Valentina

Imperial College Business School, Imperial College London, South Kensington Campus London, e-mail: v.raponi13@imperial.ac.uk

ne della distribuzione asimmetrica di Laplace multivariata. La metodologia adottata permette di modellare simultaneamente i quantili condizionati e di studiare la dipendenza tra diversi asset finanziari. Successivamente, introduciamo un approccio innovativo per la costruzione di un portafoglio finanziario la cui distribuzione risulta essere una distribuzione asimmetrica di Laplace univariata. L'analisi empirica che verrà effettuata in futuro analizzerà tre principali indici azionari e mostrerà la rilevanza applicativa della stima del VaR e dell'ES in un contesto multivariato.

Key words: Quantile Regression, Multiple quantiles, Multivariate Asymmetric Laplace Distribution, CAViaR, Value at Risk, Expected Shortfall

1 Introduction

The events of the ongoing credit crisis and past financial crises have emphasized the necessity for appropriate risk measures. The use of quantitative risk measures has become an essential management tool providing advice and support for asset management decisions. The most widely used risk measure is Value at Risk (VaR) that measures the maximum loss in which a financial operator can incur over a defined time horizon and for a given confidence level. However, it has a number of drawbacks [2]: it lacks of tail-sensitivity, thus it does not warn about the size of the losses that occur with a probability lower than the predetermined confidence level and, it is not a coherent risk measure [2]. Accordingly, as a remedy for the deficiencies of VaR, [1] and [7] introduced the Expected Shortfall (ES) defined as the conditional expectation of exceedances beyond VaR. Unlike VaR, the ES enjoys interesting properties: it is a coherent risk measure and, by taking into consideration magnitude and occurrence of extreme negative returns that drop below VaR, it offers information on the heaviness of the tails of the loss distribution.

The main goal of the present paper is to extend the work of [8] to a multiple linear quantile regression framework for VaR and ES joint forecasting. We exploit the research effort of [6] where a generalization of the univariate inferential approach based on the Asymmetric Laplace (AL) distribution to a multivariate framework has been proposed by using a reparametrization of the Multivariate Asymmetric Laplace (MAL) distribution presented by [5]. This approach allows for quantile-specific determinants of different responses and by jointly modelling them, one can borrow information across responses and conduct joint inference at the marginal quantiles level, revealing some possible underlying truth that can not be detected by univariate

models. From a practical perspective, this methodology provides a deep investigation of the interdependences among multiple financial assets at different quantile levels. To capture the salient features of financial time series, i.e. the stylized facts, we model the marginal conditional quantiles using an autoregressive specification: the Conditional Autoregressive VaR (CAViaR) of [3]. For the ES component, following the work of [8], we let the AL marginals scale parameters to be time-varying such as to produce an estimate of the time-varying conditional ES. In this way, we estimate the ES as a by-product of VaR. The inferential method is implemented by using the EM algorithm. Subsequently, we determined analytically that all linear combinations of MAL components have univariate AL densities. Hence, it is possible to build a portfolio whose returns are AL distributed where its parameters are a function of the MAL parameters and the allocation weights. It is our intention to provide investors and asset managers operative tools for portfolio construction and risk evaluation, disclosing risk factors that should be taken into account by timely and adequate re-allocations of assets. In order to assess the forecasting performance of the suggested risk measurement procedure we carry out a set of targeted tests. To assess VaR and ES predictions jointly, we use the score functions of [4] and [8], respectively.

2 Methodology

Following the work of [6], we introduce the joint quantile regression framework and the MAL distribution of [5]. Let $\mathbf{Y}_t = [Y_{t1}, Y_{t2}, \dots, Y_{tp}]'$ be a p -variate response variable and assume that the τ_j -quantile of each of the j -th component of \mathbf{Y}_t , conditional on the information set available at time $t - 1$, \mathcal{F}_{t-1} , is $Q_{Y_{tj}}(\tau_j | \mathcal{F}_{t-1})$ for $j = 1, 2, \dots, p$ and $t = 1, 2, \dots, T$. Our objective is to provide joint estimation of the p marginal conditional quantiles of $\mathbf{Y}_t \in \mathcal{R}^p$ in a multivariate framework. For a given vector $\boldsymbol{\tau} = [\tau_1, \tau_2, \dots, \tau_p]'$ we consider the model

$$\mathbf{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, 2, \dots, T \quad (1)$$

where $\boldsymbol{\varepsilon}_t$ denotes a $p \times 1$ vector of error terms with (univariate) component-wise quantiles (at fixed levels τ_1, \dots, τ_p , respectively) equal to zero. This assumption implies $\boldsymbol{\mu}_t = Q_{\mathbf{Y}_t}(\boldsymbol{\tau} | \mathcal{F}_{t-1})$ a $p \times 1$ vector whose generic element is $Q_{Y_{tj}}(\tau_j | \mathcal{F}_{t-1})$.

For model (1), we consider the following $\text{MAL}_p(\boldsymbol{\mu}_t, \mathbf{D}_t \boldsymbol{\xi}, \mathbf{D}_t \tilde{\boldsymbol{\Sigma}} \mathbf{D}_t)$ distribution with density function:

$$f_Y(\mathbf{y}_t | \boldsymbol{\mu}_t, \mathbf{D}_t \tilde{\boldsymbol{\xi}}, \mathbf{D}_t \tilde{\boldsymbol{\Sigma}} \mathbf{D}_t) = \frac{2 \exp \left\{ (\mathbf{y}_t - \boldsymbol{\mu}_t)' \mathbf{D}_t^{-1} \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\xi}} \right\}}{(2\pi)^{p/2} |\mathbf{D}_t \tilde{\boldsymbol{\Sigma}} \mathbf{D}_t|^{1/2}} \left(\frac{\tilde{m}_t}{2 + \tilde{d}} \right)^{v/2} K_v \left(\sqrt{(2 + \tilde{d}) \tilde{m}_t} \right)$$

where $\boldsymbol{\mu}_t$ is the location parameter vector, $\mathbf{D}_t \tilde{\boldsymbol{\xi}} \in \mathcal{R}^p$ is the scale (or skew) parameter, with $\mathbf{D}_t = \text{diag}[\delta_{t1}, \delta_{t2}, \dots, \delta_{tp}]$, $\delta_{tj} > 0$ and $\tilde{\boldsymbol{\xi}} = [\tilde{\xi}_1, \tilde{\xi}_2, \dots, \tilde{\xi}_p]'$, having generic element $\tilde{\xi}_j = \frac{1-2\tau_j}{\tau_j(1-\tau_j)}$. $\tilde{\boldsymbol{\Sigma}}$ is a $p \times p$ positive definite matrix such that $\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{\Lambda}} \boldsymbol{\Psi} \tilde{\boldsymbol{\Lambda}}$, with $\boldsymbol{\Psi}$ being a correlation matrix and $\tilde{\boldsymbol{\Lambda}} = \text{diag}[\tilde{\sigma}_1, \tilde{\sigma}_1, \dots, \tilde{\sigma}_p]$, with $\tilde{\sigma}_j^2 = \frac{2}{\tau_j(1-\tau_j)}$, $j = 1, \dots, p$. Moreover, $\tilde{m}_t = (\mathbf{y} - \boldsymbol{\mu}_t)' (\mathbf{D}_t \tilde{\boldsymbol{\Sigma}} \mathbf{D}_t)^{-1} (\mathbf{y} - \boldsymbol{\mu}_t)$, $\tilde{d} = \tilde{\boldsymbol{\xi}}' \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\xi}}$, and $K_v(\cdot)$ denotes the modified Bessel function of the third kind with index parameter $v = (2 - p)/2$. Under these conditions we have that, if $\mathbf{Y}_t \sim \text{MAL}_p(\boldsymbol{\mu}_t, \mathbf{D}_t \tilde{\boldsymbol{\xi}}, \mathbf{D}_t \tilde{\boldsymbol{\Sigma}} \mathbf{D}_t)$, then $\mathbb{P}(Y_{tj} < \mu_{tj}) = \tau_j$ and $Y_{tj} \sim \text{AL}(\mu_{tj}, \tau_j, \delta_{tj})$ for $j = 1, 2, \dots, p$. With respect to the MAL density employed in [6], here we let the diagonal matrix \mathbf{D}_t , be time-varying i.e. each δ_{tj} represents the time-varying scale parameter of the marginal AL distribution of Y_{tj} , for every $j = 1, 2, \dots, p$. This generalization allows us to jointly estimate the p marginal conditional VaR and ES of $\mathbf{Y}_t \in \mathcal{R}^p$.

Throughout the paper we assume that $Q_{Y_{tj}}(\tau_j | \mathcal{F}_{t-1})$ can be modelled as a function of its past values, $Q_{Y_{t-1j}}(\tau_j | \mathcal{F}_{t-2})$ using a CAViaR representation proposed in [3]. For the ES component, following the work of [8], we can express it in terms of the conditional scale matrix \mathbf{D}_t of the MAL density. Each marginal of the MAL density is AL distributed and this justifies that the ES can be expressed in terms of the conditional scale parameter δ_{tj} . We consider a time-varying scale parameter of the AL density of the type:

$$ES_{tj} = \mathbb{E}[Y_{tj}] - \frac{\delta_{tj}}{\tau_j} \quad (2)$$

where

$$\delta_{tj} = \mathbb{E}[(Y_{tj} - Q_{Y_{tj}})(\tau_j - \mathbf{1}_{(y_t < Q_{Y_{tj}})})] \quad (3)$$

In the light of the outcome of [8] and supposing that \mathbf{Y}_t has zero mean, we model the ES as the product of the quantile and a multiplicative factor γ_j . In this way, the dynamics of VaR and of ES is the same. To avoid crossing, we express the ES using an exponential function of the unconstrained parameter γ_j .

$$ES_{tj} = (1 + e^{\gamma_j}) Q_{Y_{tj}}(\tau_j), \quad t = 1, 2, \dots, T, \quad j = 1, 2, \dots, p. \quad (4)$$

The representation in (4) provides a simple and parsimonious approach to enable simultaneous VaR and ES estimation in a semiparametric framework.

In order to build a portfolio from the marginal distributions of the MAL we show that all linear combinations of the components of the $\text{MAL}_p(\mu_t, \mathbf{D}_t \tilde{\xi}, \mathbf{D}_t \tilde{\Sigma} \mathbf{D}_t)$ density are AL distributed. One can exploit this property to construct an affine combination of the MAL components with scalars $b_j \in \mathcal{R}$ for $j = 1, \dots, p$ such that $\sum_{d=1}^p b_d = 1$.

3 Empirical Study

To compare the results with those of [8], and estimate VAR and ES measures in the next future we will perform an empirical analysis on the weekly returns of the FTSE 100, Nikkei 225, and S&P 500 stock indices from 11/03/1988 to 02/11/2018 and using

4 Conclusion

In this paper we generalize the Multivariate Asymmetric Laplace joint quantile regression approach considered in [6], by modelling the marginal quantiles with the Conditional Autoregressive Value at Risk (CAViaR) structure proposed by [3] and the marginal scale parameters with a time-varying specification. In this way we are able to capture the dynamic nature of the data and estimate a time-varying conditional ES jointly with VaR following the approach of [8] for each component of the multivariate response taking into account for the possible correlation among marginals.

References

- [1] Acerbi, C. and Tasche, D. [2002], ‘On the coherence of expected shortfall’, *Journal of Banking & Finance* **26**(7), 1487–1503.
- [2] Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D. [1999], ‘Coherent measures of risk’, *Mathematical finance* **9**(3), 203–228.
- [3] Engle, R. F. and Manganelli, S. [2004], ‘CAViaR: conditional autoregressive Value at Risk by regression quantiles’, *Journal of Business & Economic Statistics* **22**(4), 367–381.
- [4] Fissler, T., Ziegel, J. F. and Gneiting, T. [2015], ‘Expected Shortfall is jointly elicitable with Value at Risk - Implications for backtesting’, *arXiv preprint arXiv:1507.00244*.

- [5] Kotz, S., Kozubowski, T. and Podgorski, K. [2012], *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*, Springer Science & Business Media.
- [6] Petrella, L. and Raponi, V. [2019], 'Joint estimation of conditional quantiles in multivariate linear regression models. An application to financial distress', *Journal of Multivariate Analysis* **173**, 70–84.
- [7] Rockafellar, R. T., Uryasev, S. et al. [2000], 'Optimization of conditional Value at Risk', *Journal of risk* **2**, 21–42.
- [8] Taylor, J. W. [2019], 'Forecasting Value at Risk and Expected Shortfall using a semiparametric approach based on the asymmetric Laplace distribution', *Journal of Business & Economic Statistics* **37**, 121–133.

Approximate Bayesian Computation methods to model Multistage Carcinogenesis

Metodi di Approximate Bayesian Computation per modellare la Cancerogenesi Multistadiale

Consuelo R. Nava, Cinzia Carota, Jordy Bollon, Corrado Magnani, Francesco Barone-Adesi

Abstract A direct modelling of Multistage Carcinogenesis (MC), avoiding mathematical approximations, is here proposed. We take advantage of Approximate Bayesian Computation methods to estimate MC unknown parameters of interest. A simulation of a fictitious cohort of people exposed to a carcinogen is proposed. We show performances of our approach with and without the use of a semi-automatic ABC selection of summary statistics.

Abstract *Si propone una modellizzazione diretta della cancerogenesi multistadiale (MC), evitando approssimazioni matematiche. Metodi di Approximate Bayesian Computation vengono utilizzati per stimare i parametri di interesse della MC. Si simula una coorte fittizia di persone esposte a un agente cancerogeno. Si mostrano le performance del nostro approccio con e senza l'uso di tecniche di selezione semiautomatica delle statistiche descrittive.*

Key words: Approximate Bayesian Computation; semi-automatic selection; rejection algorithm; Multistage Carcinogenesis.

Consuelo R. Nava

Università della Valle d'Aosta, Dipartimento di Economia e Scienze Politiche, Aosta (Italy) e-mail: c.nava@univda.it

Cinzia Carota

Università degli Studi di Torino, Dipartimento di Economia e Statistica Cogneetti de Martiis, Torino (Italy) e-mail: cinzia.carota@unito.it

Jordy Bollon

Università del Piemonte Orientale, Dipartimento di Medicina Traslazionale, Novara (Italy) e-mail: jbollon94@gmail.com

Corrado Magnani

Università del Piemonte Orientale, Dipartimento di Medicina Traslazionale, Novara (Italy) e-mail: corrado.magnani@uniupo.it

Francesco Barone-Adesi

Università del Piemonte Orientale, Dipartimento di Scienze del Farmaco, Novara (Italy) e-mail: francesco.baroneadesi@uniupo.it

1 Introduction

The theory of multistage carcinogenesis (MC) assumes that the transformation of a normal cell into a neoplastic one does not take place in a single step, but rather consists of a multi-stage process [17]. In each stage, normal cells undergo a sequence of genetic mutations which gradually cause the acquisition of tumor cell characteristics [2]. A time-homogeneous birth process governs the transition probability from the i -th to the $i + 1$ -th stage at time t , with $t = 0, \dots, T$ [1]. Approximated formulas of MC are available to predict cancer rates at different times [16, 3], avoiding algebraically cumbersome computations resulting from the system of stochastic differential equations which model MC. Accurate predictions of cancer risk rates are useful [18] to plan health surveillance programs for carcinogenic agents. However, epidemiological studies using MC are presently limited, and they usually rely only on approximated formulas [12, 5, 21].

Different authors pointed out that in some situations the use of approximated formulas, indeed, can lead to an overestimation of hazard rates [16]. Hence, it would be desirable to use the “exact” MC model to evaluate the evolution of cancer risk with the age and, eventually, with long-term carcinogenic exposure. However, due to its high complexity, MC does not allow to define a likelihood function. A possible solution is represented by Bayesian methods, now increasingly used in population genetics [13]. Specifically, Approximate Bayesian Computation (ABC) methods [14, 19] compare observed data with simulated data not through the likelihood function – assumed to be unavailable – but rather with selected summary statistics, such as means, hazard ratios (etc.), obtained through simulations from the same original model – assumed to be known [11, 13]. Even if recent ABC applications can be found in population genetics [9, 23, 22], infectious disease models [15], and systems biology [20, 26], its use in epidemiology is still limited [24, 27, 6].

As an alternative to approximated formulas [5], we propose ABC methods to model MC and to estimate its unknown parameters θ of interest. We propose a suitable Rejection algorithm [25] enriched with a semi-automatic variable selection method [10]. A code in R has been developed to model MC to estimate θ : the transition rates among the different stages (which can vary during the exposure to a carcinogen) and the elimination rates of the carcinogen from the organism.

To this aim, the article is structured as follows. In Section 2, we present ABC techniques to illustrate how ABC can be integrated to model MC and how summary statistics can be selected. In Section 3, we propose a simulation example which mimics a cohort of subjects exposed to a carcinogen and we present some preliminary results. Section 4 concludes the article with some remarks and suggesting future research.

2 Methodology

In epidemiology and population genetics, limits in the use of frequentist model-based inference arise due to the necessity to include prior information and to explicit the likelihood function.

The first issue could be overcome taking advantage of Bayesian methods. Indeed, estimates of θ given observed data \mathcal{D} are obtained sampling from the posterior distribution $\pi(\theta|\mathcal{D})$, proportional to $L(\mathcal{D}|\theta)\pi(\theta)$, respectively the likelihood function and the prior distribution of θ .

The second issue could be overcome using ABC methods as a rejection technique [19, 25, 23] to compute the likelihood function via simulation. Let's assume that an underlying stochastic process \mathcal{M} given θ generates \mathcal{D} on which k summary statistics of the data $\mathbf{S} = \{S_1, \dots, S_k\}$ are defined. ABC, in its simplest form, “proposes” a (pseudo)-randomly drawn parameter value $\theta^{(n)}$ from $\pi(\theta)$ with $n = 1, \dots, N$. Hence, simulated artificial data $\mathcal{D}^{(n)}$ are generated from \mathcal{M} given $\theta^{(n)}$. $\theta^{(n)}$ are accepted only if $\mathcal{D}^{(n)}$ is “similar enough” to \mathcal{D} and used to approximate $\pi(\theta|\mathcal{D})$ with the ABC posterior, $\pi(\theta|\mathcal{D}^{(n)})$. To this aim, an acceptance criterion of $\theta^{(n)}$ might be based on $\rho(\mathcal{D}, \mathcal{D}^{(n)}) < \varepsilon$, where $\rho(\cdot, \cdot)$ is a suitable metric and $\varepsilon > 0$ is a tolerance level. Given the high-dimensional data generated in our case, the acceptance criterion should be substituted by $\rho(\mathbf{S}, \mathbf{S}^{(n)}) < \varepsilon$ with $\mathbf{S}^{(n)} = \{S_1^{(n)}, \dots, S_k^{(n)}\}$, a small number of suitable summary statistics defined on $\mathcal{D}^{(n)}$. In such a way, we combine both the computational convenience of using \mathbf{S} to approximate $\pi(\theta|\mathbf{S}^{(n)})$ instead of $\pi(\theta|\mathcal{D}^{(n)})$, and main Bayesian inference advantages in epidemiology [8].

Rejection-sampling method [19] needs a small number k of suitable summary statistics [7] to avoid a low acceptance rate or a distorting increment of ε [4] (small values of ε allow an approximately calibrated ABC [14, 10]). The complexity of MC and the long follow-up of patients do not accommodate this requirement, making the selection of \mathbf{S} difficult and/or reducing estimation accuracy. Thus, based on a regression adjustment – namely, the robust semi-automatic ABC projection technique [10] – we added an extra stage to the algorithm in [19] in order to overcome this issue and to derive summary statistics within our ABC for MC.

The semi-automatic ABC consist of the following steps: (i) run a pilot ABC based on summary statistics chosen subjectively to identify a region of non-negligible posterior mass, i.e. a training region to simulate parameter values. This is a suitable step when uninformative or improper $\pi(\theta)$ are considered; (ii) simulate sets of parameter values $\theta^{(n)}$ from the prior truncated to the training region and generate artificial data $\mathcal{D}^{(n)}$; (iii) use $\theta^{(n)}$ and $\mathcal{D}^{(n)}$ to estimate summary statistics fitting regressions; (iv) select the best model, according to model selection criteria (as BIC), and run ABC with selected summary statistics.

Regressions of step (iii) are the linear ones [10] which have as dependent variables the simulated values of the i^{th} parameter, $\theta_i^{(1)}, \dots, \theta_i^{(n)}, \dots, \theta_i^{(N)}$, with $i = 1, 2, 3$. A vector-valued function of (non-linear) transformations of the input statistics of the artificial data, $f(\mathbf{S}^{(n)}) = [\mathbf{S}^{(n)}, \mathbf{S}^{(n)2}, \mathbf{S}^{(n)3}, \mathbf{S}^{(n)4}]$ – here all first, second, third and fourth powers of individual data point – represents the set of explanatory variables.

The following model is fitted using least squares

$$\theta_i = \mathbb{E}(\theta_i|\mathcal{D}) + \epsilon_i = \beta_{i0} + \beta_i f(\mathbf{S}) + \epsilon_i \quad \forall i = 1, 2, 3$$

where ϵ_i is a white noise error. The fitted function $\hat{\beta}_{i0} + \hat{\beta}_i f(\mathbf{S})$ is an estimate of $\mathbb{E}(\theta_i|\mathcal{D})$. Neglecting the constant, the i^{th} summary statistic for ABC is $\hat{\beta}_i f(\mathbf{S})$. Note that the input statistics, \mathbf{S} , could include raw data and (non-linear) transformations.

In general, we assume J stages in MC, denoted with E_j , with $j = 1, \dots, J$, and a constant transition rate for the cell to go from the state E_j to the state E_{j+1} , $E_j \rightarrow E_{j+1}$ [16]. Here, r_0 is the transition rate $E_1 \rightarrow E_2$, if the individual is not exposed to a carcinogen. We assume that r_0 represents also the transition rates across all the other stages, $E_j \rightarrow E_{j+1}$, with $j = 2, \dots, J-1$. Thus, besides the E_1 , transition rates are assumed to be constant (r_0). $r_1 = a \cdot r_0$ denotes the accelerated cell transition rate $E_1 \rightarrow E_2$ which is observed during the exposure to a carcinogen in E_1 , while λ represents the clearance of slowly eliminated carcinogens such as asbestos [3, 5]). Hence, $\theta = \{r_0, r_1, \lambda\}$ is the vector of unknown MC parameters of interest as proposed in (1). Given that a likelihood function cannot be defined to model MC, ABC methods allow to approximate the posterior distribution $\pi(\theta|\mathcal{D})$.

$$\begin{array}{l} \boxed{E_1} \xrightarrow{r_0} \boxed{E_2} \xrightarrow{r_0} \dots \xrightarrow{r_0} \boxed{E_j} \xrightarrow{r_0} \dots \xrightarrow{r_0} \boxed{E_J} \quad \text{no carcinogen exposure} \\ \boxed{E_1} \xrightarrow{r_1, \lambda} \boxed{E_2} \xrightarrow{r_0} \dots \xrightarrow{r_0} \boxed{E_j} \xrightarrow{r_0} \dots \xrightarrow{r_0} \boxed{E_J} \quad \text{carcinogen exposure} \end{array} \quad (1)$$

Under MC and assuming constant clearance λ of the internal dose, d_i , of the carcinogen overtime, the transition rate, r_t , at time t is:

$$\log(r_t) = \alpha + \beta \sum_{i=0}^t (d_i e^{-\lambda(t-i)}).$$

3 Simulations and preliminary results

We describe a model \mathcal{M} for carcinogen exposure to show how the proposed ABC methods accurately estimate θ and to construct a general approach to deal with MC. We simulate a fictitious cohort of 5,000 subjects, each of which was observed for a time $T = 100$, where $t = 0$ is the year of birth. A cancer develops if at least one cell reaches the last stage. We assume, without lack of generality, $J = 4$. The carcinogen exposure, consecutive or not, between 15 and 64 years old, could mimic, for instance, the asbestos exposure in an occupational setting [3]. We set $\theta = \{7 \cdot 10^{-6}, 7 \cdot 10^{-5}, 0.2\}$ and we assume that 50% of workers were exposed to asbestos. Hazard ratios for each year are computed as summary statistics and used comparatively with the semi-automatic ABC selection method given $f(\mathbf{S})$. We run 200,000 simulations with $\varepsilon = 0.005$, and uninformative priors $\pi(\theta_i) = \text{Unif}(0, 1) \quad \forall i = 1, 2, 3$, given an own elaborated R code for MC which also recalls EasyABC and abctools packages. Table 1

shows the obtained ABC preliminary estimates of θ from the approximated posterior means, with and without the semi-automatic ABC selection. The former approach results to be more accurate and closer to the original values of the parameters of interest than the one with arbitrarily selected summary statistics (hazard ratio for $t = \{35, 40, 85, 95\}$). Estimations with multiple combinations of summary statistics are carried out. No meaningful improvement of the ABC performance with respect to the one here proposed were obtained.

Table 1 Posterior means approximated with ABC rejection. Standard errors are in parentheses. Summary statistics without semi-automatic selection are hazard ratio of selected years.

Semi-automatic selection	r_0	r_1	λ
No	$6.14 \cdot 10^{-6}$ ($9 \cdot 10^{-7}$)	$6.17 \cdot 10^{-5}$ ($1.84 \cdot 10^{-5}$)	0.114 (0.078)
Yes	$7.2 \cdot 10^{-6}$ ($7 \cdot 10^{-7}$)	$6.64 \cdot 10^{-5}$ ($1.45 \cdot 10^{-5}$)	0.216 (0.078)

4 Conclusion

We show that part of the appeal of the proposed ABC approach is its flexibility. We are planning to apply the proposed methodology to a real cohort of workers exposed to asbestos to predict future mesothelioma rates. The here proposed methodology can be easily implemented to any carcinogen exposure under MC. The code written in R is general enough to accommodate other epidemiological assumptions, such as the asbestos clearance. Future research will be aimed to extend this approach to include more sophisticated ABC methods (MCMC or sequential ABC).

References

1. Armitage, P.: Multistage models of carcinogenesis. *Environ Health Perspect.* 63, 195-201 (1985)
2. Armitage, P., Doll, R.: The age distribution of cancer and multi-stage theory of carcinogenesis. *Br. J. Cancer.* 8.1, 1-12 (1954)
3. Barone-Adesi, F., Ferrante, D., Bertolotti, M., Todesco, A., Mirabelli, D., Terracini, B., Magnani, C.: Long-term mortality from pleural and peritoneal cancer after exposure to asbestos: Possible role of asbestos clearance. *Int. J. Cancer.* 123.4, 912-916 (2008)
4. Beaumont, M.A., Zhang, W., Balding, D.J.: Approximate Bayesian computation in population genetics. *Genetics.* 162.4, 2025-2035 (2002)
5. Berry, G.: Prediction of mesothelioma, lung cancer, and asbestosis in former Wittenoom asbestos workers. *Occup. Environ. Med.* 48.12, 793-802 (1991)
6. Dehideniya, M.B., Drovandi, C.C., McGree, J.M.: Optimal Bayesian design for discriminating between models with intractable likelihoods in epidemiology. *Comput. Stat. Data Anal.* 124, 277-297 (2018)
7. Didelot, X., Everitt, R.G., Johansen, A.M., Lawson, D.J.: Likelihood-free estimation of model evidence. *Bayesian Anal.* 6.1, 49-76 (2011)

8. Dunson, D.B.: Commentary: practical advantages of Bayesian analysis of epidemiologic data. *Am. J. Epidemiol.* 153.12, 1222-1226 (2001)
9. Fagundes, N.J.R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F.M., Bonatto, S.L., Excoffier, L.: Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci.* 104.45, 17614-17619 (2007)
10. Fearnhead, P., Prangle, D.: Constructing summary statistics for approximate Bayesian computation: Semi-automatic ABC (with discussion). *J R Stat Soc Series B Stat Methodol.* 74.3, 419-474 (2012)
11. Grelaud, A., Marin, J.M., Robert, C., Rodolphe, F., Tally, F.: Likelihood-free methods for model choice in Gibbs random fields. *Bayesian Anal.* 3, 427-442 (2009)
12. Magnani, C., Ferrante, D., Barone-Adesi, F., Bertolotti, M., Todesco, A., Mirabelli, D., Terracini, B.: Cancer risk after cessation of asbestos exposure: a cohort study of Italian asbestos cement workers. *Occup. Environ. Med.* 65.3, 164-170 (2008)
13. Marjoram, P., Tavaré, S.: Modern computational approaches for analysing molecular genetic variation data. *Nat. Rev. Genet.* 7.10, 759-770 (2006)
14. Marin, J.M., Pudlo, P., Robert, C.P., Ryder, R.J.: Approximate Bayesian computational methods. *Stat. Comput.* 22.6, 1167-1180 (2012)
15. McKinley, T., Cook, A.R., Deardon, R.: Inference in epidemic models without likelihoods. *Int. J. Biostat.* 5.1 (2009)
16. Moolgavkar, S.H.: The multistage theory of carcinogenesis and the age distribution of cancer in man. *JNCI.* 61.1, 49-52 (1978)
17. Nordling, C.O.: A new theory on the cancer inducing mechanism. *Br J Cancer.* 7.1, 68-72 (1953)
18. Peto, J., Decarli, A., La Vecchia, C., Levi, F., Negri, E.: The European mesothelioma epidemic. *Br. J. Cancer.* 79.3-4, 666-672 (1999)
19. Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W.: Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol.* 16.12, 1791-1798 (1999)
20. Ratmann, O., Jørgensen, O., Hinkley, T., Stumpf, M.P.H., Richardson, S., Wiuf C.: Using likelihood free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*. *PLoS Comput. Biol.* 3.11, 2266-2278 (2007)
21. Reid, A., de Klerk, N.H., Magnani, C., Ferrante, D., Berry, G., Musk, A.W., Merler, E.: Mesothelioma risk after 40 years since first exposure to asbestos: a pooled analysis. *Thorax.* 69.9, 843-850 (2014)
22. Saulnier, E., Gascuel, O., Alizon, S.: Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. *PLoS Comput Biol.* 13.3, (2017)
23. Sottoriva, A., Tavaré, S.: Population Genetics of Neoplasms. In: *Frontiers in Cancer Research*, pp. 31-42. Springer, New York (2016)
24. Tanaka, M.M., Francis, A.R., Luciani, F., Sisson, S.A.: Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics.* 173.3, 1511-1520 (2006)
25. Tavaré, S., Balding, D.J., Griffiths, R.C., Donnel, P.: Inferring coalescence times from DNA sequence data. *Genetics.* 145.2, 505-518 (1997)
26. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.: Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. Roy. Soc. Interface.* 6.31, 187-202 (2007)
27. Walker, D.M., Allingham, D., Lee, H.W.J., Small, M.: Parameter inference in small world network disease models with approximate Bayesian computational methods. *PHYSICA A.* 389.3, 540-548 (2010)

Co-clustering TripAdvisor data for personalized recommendations

Co-clustering di dati TripAdvisor per un sistema di raccomandazioni personalizzato

Giulia Pascali, Alessandro Casa and Giovanna Menardi

Abstract TripAdvisor is one of the largest travel websites. Among the provided services, it aids users with suggestions about attractions, accommodations, restaurants, etc., based on a wide system of reviews. In fact, users looking for suggestions shall sort through opinions posted by any kind of other users, possibly with different preferences and travelling behaviour. The aim of this work is to provide a personalized recommendation system to integrate the TripAdvisor services, based on the identification of similar travel products rated by similar users. Some alternative models of co-clustering are considered, to handle user rates in the form of ordinal data, and to account for missing values, due to the intrinsic fact that each user rates only a small subset of the considered travel products. Possible extensions are discussed to include additional information in the model, based on products and users characteristics.

Abstract *TripAdvisor è una delle più grandi piattaforme web dedicate al turismo. Tra i servizi offerti, il più noto è quello di fornire consigli di viaggio basati sulla valutazione delle attrazioni turistiche, hotel, e ristoranti, da parte degli utenti. Tuttavia, i viaggiatori alla ricerca di suggerimenti devono districarsi tra opinioni fornite da ogni tipologia di utente, con preferenze ed abitudini di viaggio anche molto diverse. L'obiettivo di questo lavoro è quello di fornire un sistema di raccomandazioni personalizzato ad integrazione di quello fornito da TripAdvisor. Si considerano alcuni modelli di co-clustering volti a gestire le valutazioni degli utenti nella forma di dati ordinali tenendo conto al tempo stesso della presenza di dati mancanti, problema intrinseco legato alla tendenza del singolo utente a recensire solo un esiguo sottoinsieme di prodotti turistici tra quelli disponibili. Vengono inoltre discusse possibili estensioni dei modelli, finalizzate a includere ulteriori informazioni riguardo alle caratteristiche dei prodotti recensiti e degli utenti.*

Key words: co-clustering, ordinal data, recommendation systems

Giulia Pascali, Alessandro Casa, Giovanna Menardi
Dipartimento di Scienze Statistiche, Università degli Studi di Padova
via C. Battisti 241, 35121, Padova; e-mail: giulia.pascali@studenti.unipd.it,
casa@stat.unipd.it, menardi@stat.unipd.it

1 Introduction

TripAdvisor¹ is one of the largest travel websites, providing users with suggestions about trip plans, attractions, accommodations, restaurants etc. While the main tool for advising is a wide system of reviews, users looking for suggestions shall sort through opinions provided by any kind of other users, possibly with different preferences and travelling behaviour. In this immense set of possible alternatives, personalized recommendations can come to the user aid and help addressing a choice.

Recommendation systems are becoming increasingly sophisticated in the e-commerce era aiming at providing users with more and more targeted and personalized advices. Especially based on machine learning techniques [5], recommendation systems establish connections among users and items based on the preferences of similar users or on product similarity, in order to predict users' choices. A statistical approach which lends itself to the purpose of creating personalized recommendations is known as *co-clustering*. Aimed at jointly identifying clusters of observations and variables i.e., in the specific framework, users and items, co-clustering can be seen as a way to combine recommendation systems based on collaborative filtering - that account for similarities among users only - and content-based techniques, considering similarities among items.

In this work we aim at providing a recommendation system to be integrated within the TripAdvisor services. Some alternative models of co-clustering are considered, to handle user rates in the form of ordinal data, and to account for the intrinsic problem of missing values, as each product is rated by a small subset of users only. Possible model extensions are discussed to include additional information in the model, e.g. based on product tags or user reviews.

After describing the data and their main features (Section 2), we introduce the Latent Block Model to perform co-clustering (Section 3) and overview some possible specifications for the problem at hand. Results on their application on TripAdvisor data are illustrated and discussed (Section 4), along with some model extensions.

2 Data Description

Data at hand have been downloaded from the web² via web scraping, and refer to all the restaurants and bars located in the province of Padova and rated on TripAdvisor. The considered time horizon spans from the 1st of August 2011 (date of the first review) to the 15th of November 2018 (date of the download).

The original data set includes 709 restaurants and 42.263 users. Each rate consists of an integer number between 1 (terrible) to 5 (excellent). A whole amount of 97555 ratings is observed, leading to a very sparse (99.7% of missing values) utility

¹ www.tripadvisor.com

² https://www.tripadvisor.it/Restaurants-g187867-Padua_Province_of_Padua_Veneto.html

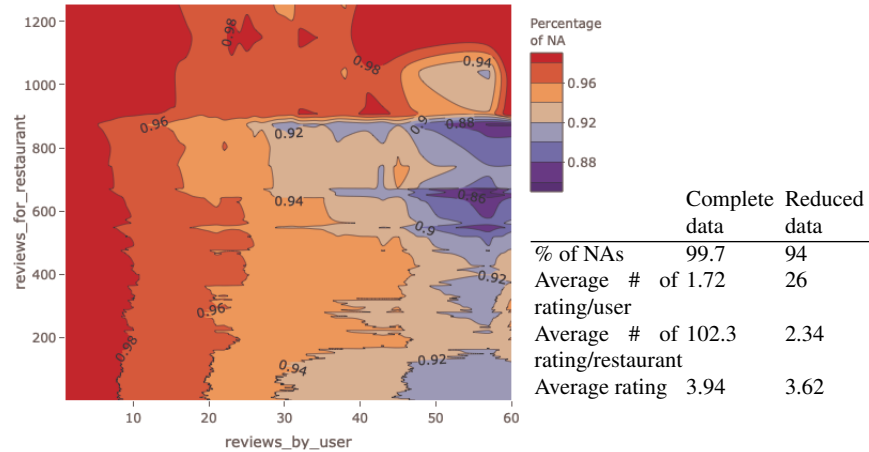


Fig. 1: Left: empirical distribution of the missing values as the number of reviews per user and the number of reviews per restaurant varies. Right: descriptive statistics of complete and reduced data.

matrix - i.e. a matrix with cells reporting the ratings and rows and columns are associated with users and restaurants respectively. While one specific strategy is required to handle such a large amount of missing values, we have proceeded with a preliminary reduction of the data in order to limit the problem as far as possible. Fig. 1 highlights how the percentage of missing data changes when removing a number of reviewing users and restaurants being reviewed. Hence, to give more weight to the most informative part of the data, a reduced utility matrix with 94% of missing data has been extracted from the original one for the analysis, including 500 restaurants (each of them rated at least by 10 users) and 45 users (each of them having rated at least 25 restaurants).

Additional information for each rate may (but not necessarily does) include the price range, the total score, a tag indicating the type of cuisine of the restaurant (e.g. italian, vegetarian, japanese etc...), and a text review.

3 Co-clustering for ordinal data

In order to build a personalized recommendation system, based on both user and item similarities, a suitable framework is represented by the so-called *co-clustering* approach. Co-clustering aims at providing a joint partition of rows and columns of a data matrix. Several approaches have been proposed in literature, following either heuristic or probabilistic perspectives [3]. Among the latter ones, the most considered one is, unarguably, the *latent block model* (LBM).

Let $\mathbf{x} = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ be the data at hand, where, in our case, each $x_{ij} \in \{1, \dots, M\}$ corresponds to the (possibly missing) rate given by the i -th user to the j -th restaurant. Two Multinomial latent variables $\mathbf{z} = (z_{ik})_{1 \leq i \leq n, 1 \leq k \leq K}$ and $\mathbf{w} = (w_{jl})_{1 \leq j \leq p, 1 \leq l \leq L}$ are introduced to describe respectively the row and the column cluster membership, where $z_{ik} = 1$ if observation i belong to row-cluster k and the same holds for w_{jl} , and K, L the number of row and column clusters. In the LBM framework the independence between \mathbf{z} and \mathbf{w} is assumed and, conditionally on \mathbf{z} and \mathbf{w} , the $n \times p$ observations x_{ij} are also independent. A general latent block model for \mathbf{x} is specified as follows:

$$p(\mathbf{x}; \theta) = \sum_{\mathbf{z} \in Z} \sum_{\mathbf{w} \in W} p(\mathbf{z}; \theta) p(\mathbf{w}; \theta) p(\mathbf{x} | \mathbf{z}, \mathbf{w}; \theta), \quad (1)$$

where Z and W are respectively the rows and columns partitions, $p(\mathbf{z}; \theta) = \prod_{ik} \rho_k^{z_{ik}}$ and $p(\mathbf{w}; \theta) = \prod_{jl} \delta_l^{w_{jl}}$; $p(\mathbf{x} | \mathbf{z}, \mathbf{w}; \theta) = \prod_{ijkl} p(x_{ij}; \alpha_{kl})^{z_{ik} w_{jl}}$ and $\theta = (\rho_k, \delta_l, \alpha_{kl})$.

Coherently with the nature of the data, we need to assume for $p(\cdot; \alpha_{kl})$ a probabilistic model which handles ordinal data. We focus on the proposal by Jacques and Biernacki [4] and on the work by Corneli et al. [2].

The former approach specifies $p(\cdot; \alpha_{kl})$ as a BOS model [1], a probability distribution designed for ordinal data and governed by the parameter $\alpha_{kl} = (\mu_{kl}, \tau_{kl})$, $\mu_{kl} \in \{1, \dots, M\}$, $\tau_{kl} \in [0, 1]$, with, respectively, the roles of location and precision.

In the latter approach [2], the generative model is based on a Gaussian latent random variable with parameters $\alpha_{kl} = (\mu_{kl}, \sigma_{kl}^2)$ sharing the same rationale as the cumulative probit model, i.e. the ordinal categories are defined by continuous intervals between pre-determined cut-points $\gamma_1, \dots, \gamma_{M-1}$. The model can manage missing data possibly not at random, via the introduction of a Bernoulli variable A_{ij} with parameter π_{kl} , describing the presence of a cell observation. Hence,

$$p(X = m; \alpha_{kl} | A_{ij} = 1) = \Phi\left(\frac{\gamma_m - \mu_{kl}}{\sigma_{kl}}\right) - \Phi\left(\frac{\gamma_{m-1} - \mu_{kl}}{\sigma_{kl}}\right) \quad (2)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a Gaussian variable. In both the models, parameters are estimated via maximum likelihood.

The reader may refer to the original works for further details.

4 Results and discussion

In the following, we analyze and compare the results of the application of the considered methods on the data introduced in Section 2. The number of co-clusters has been determined via the optimization of an information-based criterion when resorting to the approach by [2], and set to the same number when considering the work by [4], since the lack of a way for handling missing values discourages the automatic selection.

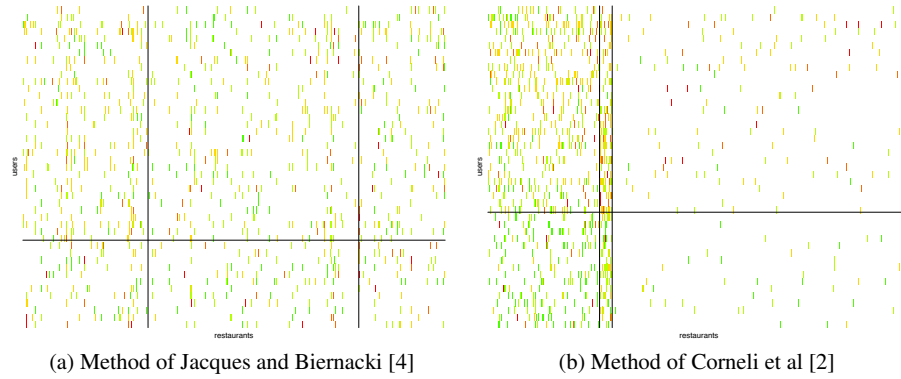


Fig. 2: Utility matrices of TripAdvisor Padova data reorganized according to the results obtained. An empty cell corresponds to a restaurant that has not been rated by the user. Different rates are associated to different colors, ranging from negative reviews (red) to positive ones (green).

In Fig. 2 the utility matrices reorganized according to the results obtained with the two approaches are illustrated. From a graphical inspection it stands out a broadly uniform distribution among the blocks obtained via the application of [4]; the intuition is confirmed by the analysis of the estimated parameters of the model. As a possible responsible for this behaviour which jeopardizes an effective interpretation of the results, we can identify the *missing at random* assumption inducing comparable frequencies of missing values in each co-cluster. Conversely, a joint inspection of Fig. 2 and Fig. 3 allows to obtain some interesting insights on the data when the method [2] is applied. Generally, it can be noticed that one row cluster groups users assigning lower rates with respect to the users in the other row cluster. The percentage of missing values varies among blocks; two co-clusters represent rarely reviewed restaurants, with average and respectively good ratings. Two further column cluster represents groups of largely reviewed restaurants, with row rates distributed coherently with the considerations above. Thus, the method [2] appears more appropriate than [4], both for the way it handles missing observations and for computational reasons.

Cluster homogeneity suggests a non-negligible informativeness of the customer-oriented recommendation system driven by the results. For example it would be reasonable to suggest, to customer in the second row cluster, the restaurants in the second column cluster where they do not have already eaten. It seems reasonable to inversely weight the suggestions with the percentage of missing values in each co-cluster, according to the rationale for which recommendations for rarely reviewed restaurants would have a greater degree of uncertainty.

Results suggest further room for improvement of the model [2], thanks to the availability of additional information about restaurants and users characteristics,

such as the price range and type of food sold. The mean of the underlying latent Gaussian random variable can be modeled via a regression function, depending on row and column specific features. This would allow us to obtain covariate-dependent block-specific means and to better characterize the induced co-clusters. As a by product, it would be possible to evaluate if covariates show a different impact on the response variable, depending on the specific block. Despite naturally sensible, the idea requires non-trivial modifications of the estimation procedure which are left for future research.

References

1. Biernacki, C., Jacques, J. (2016) Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Statistics and Computing* 26 (5), 929–943.
2. Corneli M, Bouveyron C. and Latouche P. (2019) Co-Clustering of ordinal data via latent continuous random variables and a classification EM algorithm. <https://hal.archives-ouvertes.fr/hal-01978174>. HAL Id: hal-01978174
3. Govaert G., Nadif M. (2013) Co-Clustering, Wiley-IEEE Press.
4. Jacques, J. and Biernacki, C. (2018) Model-based co-clustering for ordinal data. *Computational Statistics & Data Analysis*, 123, 101-115.
5. Ricci, F, Lior R., and Bracha S. (2015) *Recommender Systems Handbook*. Springer.

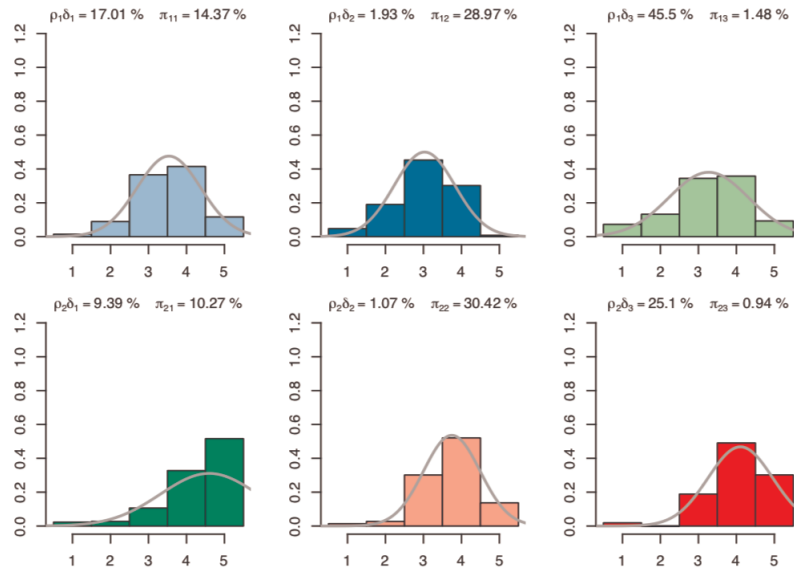


Fig. 3: Histograms of rates of the co-clusters identified by [2] with superimposed the estimated latent Gaussian variable. The estimated parameters and percentages of missing values are also reported.

Latent class analysis of endoreduplicated nuclei in confocal microscopy

Analisi di classi latenti per dati di nuclei endoreduplicati tramite microscopia confocale

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

Abstract We measured the areas of fluorescently labelled nuclei in confocal microscopy images and compared manual and semi-automated measurements based on an ImageJ plugin. Assuming such nuclear area data as a manifest variable, we try to collect K latent classes of a categorical latent variable, the ploidy level (nuclear DNA content), which can be represented as the endoreduplication index (2C for diploid, 4C for tetraploid,...etc)

Abstract Abbiamo misurato manualmente e semi-automaticamente (tramite un plugin Image-J) le dimensioni delle aree dei nuclei evidenziati con fluorescenza in immagini da microscopia confocale. Considerando le aree come variabile manifesta, applichiamo un modello a classi latenti per identificare una variabile latente che assumiamo essere il livello di ploidia ovvero il numero di appartenenza ad una classe di endoreduplicazione (2C = diploide, 4C = tetraploide, etc...).

Key words: Latent class analysis, clustering, area detection, nuclear ploidy, image analysis

1 Introduction

The topic of our research is the arbuscular mycorrhizal (AM) symbiosis, a beneficial interaction between the majority of plants and a small group of soil fungi, developing arbuscules, the structures devoted to the exchange of mineral nutrients for sugars and lipids

[1]. The interaction between symbionts involves the exchange of Carbon, Phosphorus, Nitrogen and water.

Recent evidence from our research group shed light on the activation of plant cell cycle-related mechanisms during AM colonization [2] using live microscopic observations, gene expression and flow cytometry analyses. Previous studies have highlighted endoreduplication in several plant-microbe interaction, including AM colonized roots, but the precise location of the endoreduplication events has never been defined for AM. The aim of this research is to provide more solid statistical support to the observed increase in nuclear ploidy (derived from endoreduplication events) in mycorrhizal compared to uninoculated roots, by using a clustering approach based on latent class analysis.

Endoreduplication events have been analyzed through image post-processing analyses from confocal observations to investigate the occurrence of nuclear size increases (related to increases in ploidy) in AM colonized roots of *Medicago truncatula* [2, 3].

The current presentation combines two image analysis methods, manual and customized ImageJ plugin, applied to mycorrhizal and uninoculated roots, to obtain four sample datasets to be used for the identification of latent classes using the R package poLCA.

2 Confocal microscopy images

1 cm-long root segments from both control and mycorrhizal roots were sectioned using a Vibratome into 100 μ thick slices which were then stained with DAPI, a fluorescent dye for DNA, that labels all nuclei.

These samples were then imaged under a confocal microscope, as described by [3]. In each section we identified and measured the size (area) of hundreds of nuclei. We analyzed root sections using two post-processing methods: manual measurement, more precise and reliable, but time consuming, and a customized plugin in Fiji ImageJ environment, faster but more error prone. By combining the two approaches we obtained the required datasets to be used for the detection of latent classes that could correspond to classes of nuclear ploidy.

The manual and automated areas measurements shown a frequency distribution range 15-114 μm^2 for control samples and 15-149 μm^2 for mycorrhizal root sections.

Due to the relatively small number of cells undergoing endoreduplication, simple descriptive statistics are unable to highlight significant differences between average nuclear areas (Figure 1).

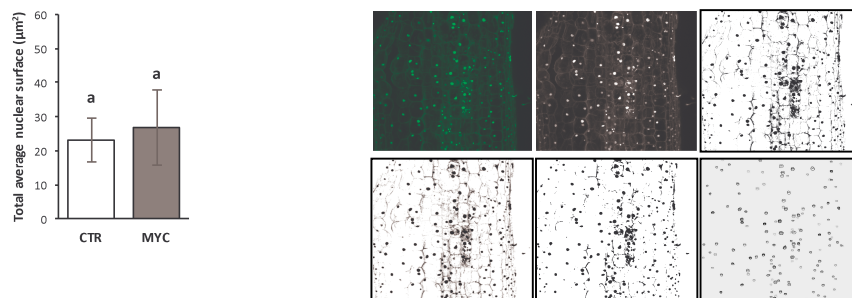


Figure 1: Detection, automated measurements and descriptive statistics

For this reason we decided to address the detection of latent classes that could correspond to the endoreduplication index. [3].

3 Latent class data analysis

We analysed four samples of nuclei areas, two deriving from manual and two from automated analysis of nuclear size in control and mycorrhizal roots.

As the poLCA package in R [4] accept only integers and identical vector of measurements we inserted missing values with each mean for each measurements in order to collect 4600 (1150 each) total spots.

Based on our previous analyses, we performed descriptive statistics with SPSS package to arbitrarily identify 8 levels of area dimensions according to 8 point division in equal classes (percentiles) for the frequency distributions.

To identify latent classes we perform poLCA code in R searching for optimal clustering according to information criteria [5]. We then compared control and mycorrhizal groups to check for a similar classification with poLCA.

The area is the manifest variable and ploidy level is assumed to be the latent variable; class membership with only one manifest variable without covariates can be computed in poLCA:

$$f(y_i; \pi_r) = \prod_{k=1}^K (\pi_{rk})^{y_{ik}}$$

And the density function

$$P(y_i | \pi, p) = \sum_{r=1}^R pr \prod_{k=1}^K (\pi_{rk})^{y_{ik}}$$

Where y_i are the areas, and π_{rk} the class conditional probability that an observation in class $r=1\dots R$ produces the k outcome of the area manifest variable and $\sum_r pr = 1$ is the weighted sum of probability for each k classes.

Then the poLCA use the EM algorithm to maximize the log likelihood function to estimate latent class model:

$$\ln L = \sum_{i=1}^N \ln pr \prod_{k=1}^K (\pi_{rk})^{y_{ik}}$$

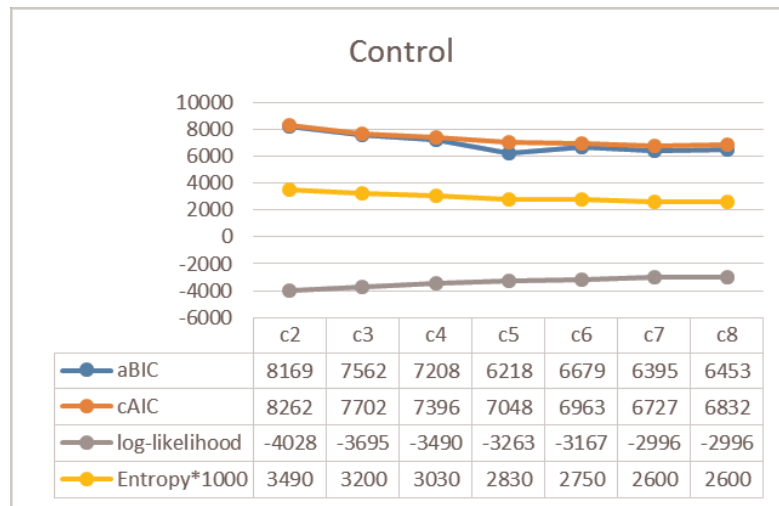
After finding latent classes we apply the information criteria: adjusted Bayesian information criterion aBIC, consistent Aikake information Criterion cAIC, log likelihood, and Entropy for finding an optima model both for control and mycorrhizal samples. For every class from $k=2$ to $k=8$ we computed the R code and the information criteria.

Models calculation:

```
library(poLCA)
data(Control or Myc)
f<- cbind(CtrManual, CtrPlugin) ~1; C_K<-poLCA(f, Control, nclass=K)
```

Information criteria:

```
aBIC<- -2*cK$llik+cK$npar*log((cK$N+2)/24)
cAIC<- -2*cK$llik+cK$npar*log(cK$N)+1
log-likelihood<-cK$llik
Entropy<-poLCA.entropy(cK)
```



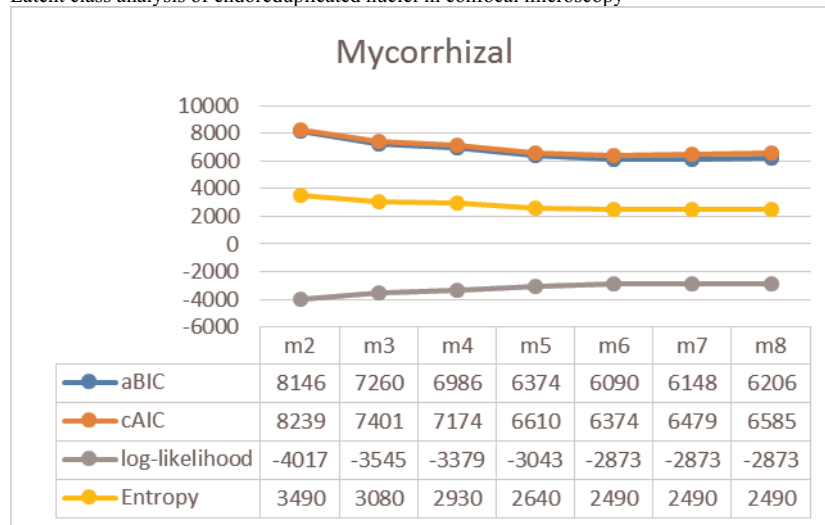


Figure 2: Optimal poLCA classes in control and mycorrhizal samples according to information criteria

Searching for lowest aBIC and cAIC, high Entropy and Log-Likelihood to individuate the optimal classes both for Control and Mycorrhizal samples we found in Control minimum aBIC in $k=5$ and cAIC in $k=7$ suggesting no accordance. In Mycorrhizal sample both aBIC and cAIC suggest optimal class for $k=6$.

Considering that cAIC is more reliable for small samples, we put more focus on the aBIC criterion. This suggests 5 optimal classes for Control sample and 6 optimal classes for Mycorrhizal sample. This result is in agreement with previous evaluations based on flow cytometry, suggesting the existence of 5 ploidy classes in control and 8 ploidy classes in mycorrhizal roots [3]. We speculate that the more restrictive statistical analysis performed by poLCA may underestimate the two top nuclear classes (actually represented by a very small number of measures), but our study provides a strong evidence that image analysis-based measurements of nuclear size (manifest variable) can correlate with actual nuclear DNA content, the ploidy level (latent variable), and convincing support to our previous interpretations of the experimental results [3].

References

1. Gutjahr C, Parniske M. Cell and developmental biology of the arbuscular mycorrhiza symbiosis. *Annu. Rev. Cell Dev. Biol.* 29: 593–617.2 (2013)
2. Kondorosi E, Roudier F, Gendreau E. Plant cell-size control: growing by ploidy? *Curr. Opin. Plant Biol.* 3:488–492 (2000)
3. Carotenuto G, Volpe V., Russo G., Politi M., Sciascia I., de Almeida-Engler J., Genre A., Local endoreduplication as a feature of intracellular fungal accommodation in arbuscular mycorrhizas. *New Phitol.* doi: 10.1111/nph.15763 (2019)
4. Linzer, D.A., Lewis J.B., poLCA: An R Package for Polytomous Variable Latent Class Analysis. *J Stat Softw* 42(10), 2-29 (2011)

Sciascia I., Carotenuto G., Genre A.

5. Zhang Z., Abarda A., Contractor A.A., Wang J., Dayton C.M., Exploring heterogeneity in clinical trials with latent class analysis. *Ann Transl Med* 6(7):119 (2018) doi: 10.21037/atm.2018.01.24