UNIVERSITÀ CATTOLICA del Sacro Cuore

Università Bocconi MILANO

POLITECNICO MILANO 1863

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

UNIVERSITÀ DEGLI STUDI DI MILANO

mathesia

Smart Statistics for Smart Applications

Book of Short Papers SIS2019

SIS 2019
SMART STATISTICS FOR SMART APPLICATIONS

SIS Società Italiana di Statistica

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
Antonio Canale

Smart Statistics: concept, technology and service 17
David John Hand, Maurizio Vichi

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

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
Mirko Armillotta, Alessandra Luati and Monia Lupparelli

An extension of the censored gaussian lasso estimator 39
Un’estensione dello stimatore clgasso
Luigi Augugliaro and Gianluca Sottile and Veronica Vinciotti

A formal approach to data swapping and disclosure limitation techniques 47
Un approccio formale per tecniche di trasformazione dei dati in problemi di privacy
F. Ayed, M. Battiston and F. Camerlenghi

A new ordinary kriging predictor for histogram data in L2-Wasserstein space 55
Un nuovo predittore kriging per istogrammi nello spazio L2-Wasserstein
Antonio Balzanella and Antonio Irpino and Rosanna Verde

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
Carolina Becatti, Irene Crimaldi and Fabio Saracco

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
Cristina Bernini, Silvia Emili, Maria Rosaria Ferrante

Statistical sources for cybersecurity and measurement issues 79
Fonti statistiche per la sicurezza cibernetica e problemi di misurazione
Claudia Biancotti, Riccardo Cristadoro, Raffaele Tartaglia Polcini

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
Chiara Bocci, Leonardo Piccini and Emilia Rocco

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
Nicolas Bousquet and Walid Dabachine

Time Series Forecasting: Is there a role for neural networks? 103
Le Reti Neuronali nella Previsione di Serie Storiche
Giuseppe Bruno, Sabina Marchetti, Juri Marcucci, Diana Nicoletti

Modelling weighted signed networks......111
Modellazione di reti segnate pesate
Alberto Caimo and Isabella Gollini

Issues on Bayesian nonparametric measures of disclosure risk 119
Questioni su misure Bayesiane nonparametriche di rischio di ”disclosure”
Federico Camerlenghi, Cinzia Carota and Stefano Favaro

Hierarchies of nonparametric priors......125
Gerarchie di distributzioni iniziali nonparametriche
Federico Camerlenghi, Stefano Favaro and Lorenzo Masoero

Issues with Nonparametric Disclosure Risk Assessment......133
Questioni sull’Analisi Nonparametrica del Rischio di “Disclosure”
Federico Camerlenghi, Stefano Favaro, Zacharie Naulet and Francesca Panero

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.
Stefano Campostrini and Lucia Zanotto

Temporal sentiment analysis with distributed lag models 149
Analisi temporale del “sentiment” con modelli a lag distribuiti
Carrannante M., Mattera R., Misuraca M., Scepi G., Spano M.

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
Rosella Castellano, Marco Mancinelli and Pasquale Sarnacchiaro

Bayesian Model Comparison based on Wasserstein Distances......167
Confronto di Modelli Bayesiani tramite Distanze di Wasserstein
Marta Catalano, Antonio Lijoi and Igor Prünster

Hierarchical Clustering and Dimensionality Reduction for Big Data 173
Clustering e Riduzione Dimensionale Gerarchici per Dati di Grandi Dimensioni
Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria

ICOs success drivers: a textual and statistical analysis......181
Fattori di successo nelle ICOs: un’analisi testuale e statistica
Paola Cerchiello and Anca Mirela Toma

Small area estimators with linked data......189
Stimatori per piccole aree nel caso di dati ottenuti attraverso il record linkage
Chambers Raymond and Fabrizi Enrico and Salvati Nicola

Optimal Portfolio Selection via network theory in banking and insurance sector......197
Gian Paolo Clemente, Rosanna Grassi and Asmerilda Hitaj

Matching error(s) and quality of statistical matching in complex surveys......205
Errori di matching e qualità del matching statistico in indagini complesse
Pier Luigi Conti and Daniela Marella

Hotel search engine architecture based on online reviews’ content......213
Un motore di ricerca per gli hotel basato sulle recensioni online
Claudio Conversano, Maurizio Romano and Francesco Mola

Economic Crisis and Earnings Management: a Statistical Analysis 219
Crisi Economica e Gestione degli Utili: un’Analisi Statistica
C. Cusatelli, A.M. D’Uggento, M. Giacalone, F. Grimaldi

A Comparison of Nonparametric Bivariate Survival Functions......227
Confronto tra stimatori non-parametrici della funzione di sopravvivanza bivariata
Hongsheng Dai and Marialuisa Restaino

Predictive Algorithms in Criminal Justice......237
Algoritmi predittivi e giustizia penale
Francesco D’Alessandro

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
Domenico De Stefano and Francesco Santelli

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
Roberta De Vito and Barbara Engelhardt

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
Davide Di Cecco, Marco Di Zio and Brunero Liseo

Spherical regression with local rotations and implementation in R....................269
Regressione sferica con rotazioni locali ed implementazione in R
Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor

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
Jacopo Diquigiovanni

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
Clelia Di Serio, Luca Del Core, Eugenio Montini and Andrea Calabria

Smart Data For Smart Health....................293
Smart Data Per Smart Health
Clelia Di Serio, Ernst C. Wit, Elena Bottinelli and Roberto Buccione

Detecting and classifying moments in basketball matches using sensor tracked data........297
Una procedure per identificare e classificare momenti di gioco in pallacanestro con l'uso di dati sensori.
Tullio Facchinetti and Rodolfo Metulini and Paola Zuccolotto

Ordered response models for cyber risk 305
Modelli a risposta ordinale per la valutazione del cyber risk
Silvia Facchinetti and Claudia Tarantola

Functional data analysis-based sensitivity analysis of integrated assessment Models for climate change modelling 313
Analisi di sensitività basata sull'analisi di dati funzionali per modelli di valutazione integrata dei cambiamenti climatici
Matteo Fontana, Massimo Tavoni and Simone Vantini

Coupled Gaussian Processes for Functional Data Analysis....................319
Processi gaussiani per l'analisi dei dati funzionali
L. Fontanella, S. Fontanella, R. Ignaccolo, L. Ippoliti, P. Valentini

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
F. Fortuna, T. Di Battista and S.A. Gattone

Statistical analysis of Sylt's coastal profiles using a spatiotemporal functional model............331
Rik Gijsman, Philipp Otto, Torsten Schlurmann, Jan Visscher

Bootstrap prediction intervals for weighted TAR predictors 339
Intervalli di previsione bootstrap per previsori ponderati per modelli TAR
Francesco Giordano and Marcella Niglio

A rank graduation index to prioritise cyber risks....................347
Un indice di graduazione per assegnare livelli di priorità ai rischi informatici
Paolo Giudici and Emanuela Raffinetti

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
Paolo Giudici and Paolo Pagnottoni

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
Luca Grassetti, Ruggero Bellio, Giovanni Fonseca and Paolo Vidoni

Trajectory clustering using adaptive squared distances....................371
Clustering di triettorie attraverso distanze adattative quadratiche
Antonio Irpino

Bayesian Analysis of Privacy Attacks on GPS Trajectories 379
Analisi Bayesiana degli Attacchi alla Privacy su Traiettorie GPS
Sirio Legramanti

Data Analytics in the Insurance Industry: Market trends and lessons from a use case customer predictive modelling 387
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

BasketballAnalyzeR: the R package for basketball analytics 395
BasketballAnalyzeR: il pacchetto R per l'analisi dei dati nella pallacanestro
Marica Manisera, Marco Sandri and Paola Zuccolotto

Data Integration by Graphical Models 403
Utilizzo dei modelli grafici per l'integrazione dei dati
Daniela Marella and Paola Vicard and Vincenzina Vitale

A two-part finite mixture quantile regression model for semi-continuous longitudinal data 409
Maruotti Antonello, Merlo Luca and Petrella Lea

Multivariate change-point analysis for climate time series 415
Analisi di change-point multivariati per serie storiche climatiche
Gianluca Mastrantonio, Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio and Carlo Blasi

A divide-et-impera approach for the spatial prediction of object data over complex regions...423
Un approccio divide-et-impera per la previsione spaziale di dati oggetto su regioni complesse
Alessandra Menafoglio e Piercesare Secchi

A strategy for the matching of mobile phone signals with census data..................................427
Una strategia per l'abbinamento di segnali di telefonia mobile con dati censuari
Rodolfo Metulini and Maurizio Carpita

Risk-based analyses for non-proportional reinsurance pricing 435
Analisi Risk-based per il pricing nella riassicurazione di trattati non proporzionali
Fabio Moraldi and Nino Savelli

A Simplified Efficient and Direct Unequal Probability Resampling 441
Un semplice Ricampionamento, efficiente e diretto per campioni a probabilità variabili
Federica Nicolussi, Fulvia Mecatti and Pier Luigi Conti

Labour Law: Machine vs. Employer Powers Diritto del lavoro: Macchina vs. Poteri datoriali 449
Antonella Occhino – Michele Faioli

Domain knowledge based priors for clustering...455
Distribuzioni a priori per l'analisi di raggruppamento basate sulla conoscenza di settore
Sally Paganin

Clustering of Behavioral Spatial Trajectories in Neuropsychological Assessment 463
Analisi dei gruppi di traiettorie spaziali nella valutazione neuropsicologica
Francesco Palumbo, Antonio Cerrato, Michela Ponticorvo, Onofrio Gigliotta, Paolo Bartolomeo, Orazio Miglino

What is wrong in the debate about smart contracts...471
Smart contract e diritto: riflessioni critiche su un dualismo fuorviante
Roberto Pardolesi and Antonio Davola

Financial Transaction Data for the Nowcasting in Official Statistics..485
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.

On the examination of a criticality measure for a complex system in a forecasting perspective 493
Esame di una misura di criticità per un sistema complesso in una prospettiva previsiva
Renata Rotondi and Elisa Varini

Knowledge discovery for dynamic textual data: temporal patterns of topics and word clusters in corpora of scientific literature 501
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

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
Andrea Sciandra, Alessio Surian and Livio Finos

Classification of spatio-temporal point pattern in the presence of clutter using K-th nearest neighbour distances....517
Classificazione dei processi puntuali spazio-temporali basata sulla distanza dal K-mo vicino più vicino
Siino Marianna, Francisco J. Rodríguez-Cortés, Jorge Mateu, Giada Adelfio

Modelling properties of high-dimensional molecular systems 525
La modellazione di sistemi molecolari ad alta dimensionalità
Debora Slanzi, Valentina Mameli and Irene Poli

Non-crossing parametric quantile functions: an application to extreme temperatures 533
Il problema del crossing con funzioni quantiliche parametriche: un'applicazione alle temperature estreme
Gianluca Sottile and Paolo Frumento

A new tuning parameter selector in lasso regression....541
Un nuovo criterio di selezione per il parametro di penalizzazione nella regressione lasso
Gianluca Sottile and Vito MR Muggeo

Similarity patterns, topological information and credit scoring models 549
Strutture di similaritá, informazioni topologiche e modelli di credit scoring
Alessandro Spelta, Branka Hadji-Misheva and Paolo Giudici

Between hawks and doves: measuring central bank communication 557
Fra falchi e colombe: valutazione delle comunicazioni di Banca Centrale
Ellen Tobback, Stefano Nardelli, David Martens

New methods and data sources for the population census 561
Nuovi metodi e fonti per il censimento della popolazione
Paolo Valente

FinTech and the Search for "Smart" Regulation 569
Fintech e la ricerca di una regolamentazione "smart"
Silvia Vanon

An anisotropic model for global climate data 577
Un modello anisotropico per i dati climatici globali
Nil Venet and Alessandro Fassò

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
Maria Kelly Venezuela, Anna Crisci, Luigi D'Ambra, D'Ambra Antonello

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
Yaqiong Wang, Alessandro Fassò and Francesco Finazzi

Tempering and computational efficiency of Bayesian variable selection....599
Tempering e l'efficienza computazionale della selezione bayesiana delle variabili
Giacomo Zanella and Gareth O. Roberts

Dimensions and links for Hate Speech in the social media 607
Dimensioni e legami per i discorsi di odio nei social media
Emma Zavarrone, Guido Ferilli

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
Antonino Abbruzzo, Mauro Ferrante, Stefano De Cantis

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
Konrad Abramowicz, Alessia Pini, Lina Schelin, Sara Sj¨ostedt de Luna, Aymeric Stamm, and Simone Vantini

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
Arianna Agosto and Emanuela Raffinetti

Monitoring SDGs at territorial level: the case of Lombardy..637
Il monitoraggio degli SDGs a livello territoriale: il caso della Lombardia
Leonardo Alaimo, Livia Celardo, Filomena Maggino, Adolfo Morrone, Federico Olivieri

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
Giacomo Aletti, Marco Bellan and Alessandra Micheletti

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
Eleonora Arnone, Laura Azzimonti, Fabio Nobile, Laura M. Sangalli

A network analysis of museum preferences: the Firenzecard experience. 653
Un'analisi di rete delle preferenze museali: l'esperienza della Firenzecard
Silvia Bacci, Bruno Bertaccini, Roberto Dinelli, Antonio Giusti, and Alessandra Petrucci

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
Michela Battauz

Tree-based Functional Data Analysis for Classification and Regression.................................665
Alberi di Classificazione e Regressione per dati Funzionali
Edoardo Belli, Enrico Ragaini, Simone Vantini

PDE-regularized regression for anisotropic 669
spatial fields Regressione con regolarizzazione differenziale per campi spaziali anisotropi
Mara S. Bernardi, Michelle Carey, James O. Ramsay and Laura M. Sangalli

A Bayesian model for network flow data: an application to BikeMi trips 673
Giulia Bissoli, Celeste Principi, Gian Matteo Rinaldi, Mario Beraha and Alessandra Guglielmi

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
Alessandra R. Brazzale and Valentina Mameli

Bayesian Variable Selection for High Dimensional Logistic Regression 685
Selezione bayesiana delle variabili nel modello di regressione logistica ad alta dimensionalita
Claudio Busatto, Andrea Sottosanti and Mauro Bernardi

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
Annalisa Cadonna, Andrea Cremaschi, Alessandra Guglielmi

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.
Anna Calissano, Aasa Feragen, Simone Vantini

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
Maria Gabriella Campolo, Antonino Di Pino and Michele Limosani

Unfolding the SEcrets of LongEvity: Current Trends and future prospects (SELECT) 709
A path through morbidity, disability and mortality in Italy and Europe
Stefano Campostrini, Daniele Durante, Fabrizio Faggiano and Stefano Mazzuco

Galaxy color distribution estimation via dependent nonparametric mixtures 713
Stima della distribuzione del colore delle galassie via misture nonparametriche dipendenti
Antonio Canale, Riccardo Corradin and Bernardo Nipoti

A case for order optimal matching: a salary gap study...719
Un algoritmo di matching ottimale ordinato per un studio sulle differenze salariali
Massimo Cannas

A Prediction Method for Ordinal Consistent Partial Least Squares 725
Un Metodo di Previsione per l'Algoritmo Ordinal Consistent Partial Least Squares
Gabriele Cantaluppi and Florian Schuberth

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
Christian Capezza, Antonio Lepore, Alessandra Menafoglio, Biagio Palumbo, and Simone Vantini

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
Andrea Cappozzo, Francesca Greselin and Giancarlo Manzi

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)
Paolo Emilio Cardone

Revenue management strategies and Booking.com ghost rates: a statistical analysis 751
Strategie di revenue management e Booking.com ghost rates: un'analisi statistica
Cinzia Carota, Consuelo R. Nava, Marco Alderighi

Analysing international migration flows: a Bayesian network approach 757
Analisi dei flussi migratori internazionali attraverso l'impiego di modelli grafici
Federico Castelletti and Emanuela Furfaro

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
Fabio Centofanti, Matteo Fontana, Antonio Lepore, and Simone Vantini

Robustness and fuzzy multidimensional poverty indicators: a simulation study.....................769
Robustezza ed indicatori fuzzy multidimensionali della povertà: uno studio di simulazione
Michele Costa

Text Based Pricing Modelling: an Application to the Fashion Industry 775
Modellazione dei prezzi basata su dati testuali: un'applicazione all'industria fashion
Federico Crescenzi, Marzia Freo and Alessandra Luati

Model based clustering in group life insurance via Bayesian nonparametric mixtures 781
Raggruppamento basato sul modello nel settore assicurativo: un approccio bayesiano nonparametrico
Laura D'Angelo

Smart Tools for Academic Submission Decisions: Waiting Times Modeling 787
Strumenti "Smart" per sottoporre i manoscritti accademici: modelli per i tempi di attesa
Francesca De Battisti - Giancarlo Manzi

On the Use of Control Variables in PLS-SEM 793
Sull'Uso delle Variabili di Controllo nei PLS-SEM
Francesca De Battisti and Elena Siletti

Partial dependence with copula and financial applications 799
Dipendenza parziale con funzioni copula e applicazioni finanziarie
Giovanni De Luca, Marta Nai Ruscone and Giorgia Rivieccio

Exploring the relationship between fertility and well-being: What is smart?............................805
Esplorando la relazione tra fecondità e benessere: cosa c'è di smart?
Alessandra De Rose, Filomena Racioppi, Maria Rita Sebastiani

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
Domenico De Stefano, Vittorio Fuccella, Susanna Zaccarin

A new regression model for bounded multivariate responses..817
Un nuovo modello di regressione per risposte multivariate limitate
Agnese Maria Di Brisco, Roberto Ascari, Sonia Migliorati and Andrea Ongaro

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
Oleksandr Didkovskyi, Alessandra Menafoglio, Piercesare Secchi, Giovanni Azzone

Hidden Markov Model estimation via Particle Gibbs 829
Stima di Hidden Markov Model tramite Particle Gibbs
Pierfrancesco Alaimo Di Loro, Enrico Ciminello and Luca Tardella

A note on marginal effects in logistic regression with independent covariates 837
Una nota sugli effetti marginali nella regressione logistica con covariate indipendenti
Marco Doretti

DNA mixtures: a case study involving a Romani reference population....................843
Misture di DNA: un caso di studio riguardante una popolazione di riferimento dei Rom
Francesco Dotto, Julia Mortera and Vincenzo Pascali

Pivotal seeding for K-means based on clustering ensembles 849
Inizializzazione pivotale dell'algoritmo delle K-medie tramite raggruppamento con metodi di insieme
Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli

Optimal scoring of partially ordered data, with an application to the ranking of smart cities855
Scoring ottimale di dati parzialmente ordinati, con un'applicazione al ranking delle smart city
Marco Fattore, Alberto Arcagni, Filomena Maggino

Bounded Domain Density Estimation 861
Stima della densità non-parametrica su domini bidimensionali limitati
Federico Ferraccioli, Laura M. Sangalli and Livio Finos

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
Ferretti C., Crosato L., Cipollini F., Ganugi P.

Design of Experiments, aberration and Market Basket Analysis....................873
Pianificazione degli esperimenti, aberrazione e Market Basket Analysis
Roberto Fontana and Fabio Rapall

Generalized Procrustes Analysis for Multilingual Studies 879
Analisi Procustiana Generalizzata per studi Multilingue
Alessia Forciniti, Michelangelo Misuraca, Germana Scepi, Maria Spano

Prior specification in flexible models....................885
Specificazione delle prior in modelli flessibili
Maria Franco-Villoria, Massimo Ventrucci and Haavard Rue

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
S. T. Gaito - G. Manzi - G. Saibene - S. Salini - M. Zignani

A PARAFAC-ALS variant for fitting large data sets 895
Una variante del PARAFAC-ALS per approssimare data set di grandi dimensioni
Michele Gallo, Violetta Simonacci and Massimo Guarino

A Convex Mixture Model for Binomial Regression 901
Un modello mistura convessa per la Regressione Binomiale
Luisa Galtarossa and Antonio Canale

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
Luigi Grossi and Mauro Mussini

Are Real World Data the smart way of doing Health Analytics? 919
Real World Data: la base di una nuova ricerca clinica?
Francesca Ieva

Internet use and leisure activities: are all young people equal? 925
Internet e tempo libero: i giovani sono uguali tra loro?
Giuseppe Lamberti, Jordi Lopez Sintas and Pilar Lopez Belbeze

On a Family of Transformed Stochastic Orders 931
Su una famiglia di ordinamenti stocastici trasformati
Tommaso Lando and Lucio Bertoli-Barsotti

Bayesian stochastic search for Ising chain graph models..935
Ricerca stocastica Bayesiana per modelli grafici a catena Ising
Andrea Lazzerini · Monia Lupparelli · Francesco C. Stingo

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 basato sull'indice di capacità di processo Cpk
Antonio Lepore, Biagio Palumbo and Philippe Castagliola

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
Alessandro Magrini

Inclusive growth in European countries: a cointegration analysis 953
La crescita inclusiva nei peasi europei: un'analisi di cointegrazione
Paolo Mariani, Andrea Marletta, Alessandra Michelangeli

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
Cristilla Martelli, Laura Grassini, Adham Kahlawi, Maria Flora Salvatori, Lucia Buzzigoli

Hidden Markov Models for High Dimensional Data 965
Hidden Markov Models per dati ad alta dimensionalità
Martino, A., Guatteri, G., Paganoni, A.M.

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
Chiara Masci, Francesca Ieva, Tommaso Agasisti and Anna Maria Paganoni

Data Mining Application to Healthcare Fraud Detection: Two-Step Unsupervised Clustering Method for Outlier Detection with Administrative Databases..977
Data Minining 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

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
M. Mingione, G. Jona Lasinio, S. Martino, F. Colloca

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
Giorgio E. Montanari and Marco Doretti

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
Marco Novelli and Maroussa Zagoraiou

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
Niels Asken Lundtorp Olsen, Alessia Pini, and Simone Vantini

Educational mismatch and attitudes towards migration in Europe..1009
Disallineamento fra formazione e lavoro e atteggiamenti verso le migrazioni in Europa
Marco Guido Palladino and Emiliano Sironi

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
Matteo Pedone, Francesco C. Stingo

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
Mattia Pedrini, Sebastian Donoso, Enrico Deusebio, Nicola Donelli, Gabriele Arici, Andrea Cosentini, Paola Mosconi, Diego Ostinelli and Claudio Cocchis

Betting on football: a model to predict match outcomes 1027
Scommettere sul calcio: un nuovo modello per prevedere l'esito delle partite
Marco Petretta, Lorenzo Schiavon and Jacopo Diquigiovanni

Estimation of dynamic quantile models via the MM algorithm 1033
Stima di modelli Quantilici Dinamici con algoritmo MM
Fabrizio Poggioni, Mauro Bernardi, Lea Petrella

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
Francesco Porro and Mariangela Zenga

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
Paola Riva, Paola Sturla, Anna Calissano and Simone Vantini

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
Lorenzo Schiavon and Nicola Sartori

Looking for Efficient Methods to Collect and Geolocalise Tweets 1057
Alla ricerca di metodi efficienti per raccogliere e geolocalizzare tweet
Stephan Schlosser, Daniele Toninelli and Silvia Fabris

Principal ranking profiles...1063
Principal ranking profiles
Mariangela Sciandra, Antonella Plaia

A statistical model for voting probabilities 1069
Un modello statistico per le probabilità di voto
Rosaria Simone, Stefania Capecchi

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
Gloria Solano-Hermosilla, Fabio Micale, Vincenzo Nardelli, Julius Adewopo, Celso Gorrín González

Dealing with uncertainty in automated test assembly problems..1083
La gestione dell'incertezza nei problemi di assemblaggio automatizzato dei test
Giada Spaccapanico Proietti, Mariagiulia Matteucci and Stefania Mignani

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
Marta Spreafico, Francesca Ieva

Bayesian multiscale mixture of Gaussian kernels for density estimation 1095
Stima di densità tramite misture bayesiane multiscala di kernel gaussiani
Marco Stefanucci and Antonio Canale

Dynamic Bayesian clustering of running activities..1101
Clustering Bayesiano dinamico di attività di corsa
Mattia Stival and Mauro Bernardi

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
Valentina Tocchioni, Daniele Vignoli, Alessandra Mattei, Bruno Arpino

A Functional Data Analysis Approach to Study a Bike Sharing Mobility Network in the City of Milan 1113
Agostino Torti, Alessia Pini and Simone Vantini

Multiresolution Topological Data Analysis for Robust Activity Tracking 1119
Giovanni Trappolini, Tullia Padellini, and Pierpaolo Brutti

Semilinear regression trees..1125
Alberi di regressione semilineari
Giulia Vannucci and Anna Gottard

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
Veronica Villani, Giuliana Barbato, Elvira Romano and Paola Mercogliano

Digital Inequalities and ICT Devices: The ambiguous Role of Smartphones...........................1139
Laura Zannella, Marina Zannella

Section 4. Posters

Modelling Hedonic Price using semiparametric M-quantile regression 1147
Regressione m-quantilica semiparametrica per la modellizzazione dei prezzi edonici
Riccardo Borgoni, Antonella Carcagnì, Alessandra Michelangeli, Nicola Salvati

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
Crescenza Calculli, Alessio Pollice, Marco V. Guglielmi and Porzia Maiorano

Official statistics to support the projects of A Scuola di OpenCoesione 1159
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.

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
Andrea Gilardi, Riccardo Borgoni and Diego Zappa

Variable selection and classification by the GRID procedure 1171
Selezione e classificazione delle variabili attraverso il metodo GRID
Francesco Giordano, Soumendra Nath Lahiri and Maria Lucia Parrella

Joint VaR and ES forecasting in a multiple quantile regression framework............................1177
Stima congiunta del VaR e dell'ES attraverso la regressione quantilica multipla
Merlo Luca, Petrella Lea and Raponi Valentina

Approximate Bayesian Computation methods to model Multistage Carcinogenesis 1183
Metodi di Approximate Bayesian Computation per modellare la Cancerogenesi Multistadiale
Consuelo R. Nava, Cinzia Carota, Jordy Bollon, Corrado Magnani, Francesco Barone-Adesi

Co-clustering TripAdvisor data for personalized recommendations 1189
Co-clustering di dati TripAdvisor per un sistema di raccomandazioni personalizzato
Giulia Pascali, Alessandro Casa and Giovanna Menardi

Latent class analysis of endoreduplicated nuclei in confocal microscopy.............................1195
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

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 unitá erroneamente conteggiate (false catture). 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 é 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 uncatchable 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, $i = 1,...,k$ (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 $\mathbf{Y} = (Y_1,\ldots,Y_k)$ 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 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 $\mathbf{y} = \mathbf{0} = (0,\ldots,0)$ 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

$$p_{\mathbf{y}} = \sum_{x=0,1} p_x \prod_{i=1}^{k} p_{y_i|x}, \tag{1}$$

where $p_{y_i|x}$ indicates the conditional probability $P(Y_i = y_i|X = x)$, or as in the log-linear model notation

$$[XY_1][XY_2]\cdots[XY_k], \tag{2}$$

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, $\{\mathscr{C}_1,\ldots,\mathscr{C}_g\}$ are the maximal cliques of G, and $(\mathscr{S}_2,\ldots,\mathscr{S}_g)$ are defined as

$$\mathscr{S}_i = \mathscr{C}_i \cap \bigcup_{j=1}^{i-1} \mathscr{C}_j \qquad i = 2,\ldots,g,$$

the joint distribution can be written as the product of conditional distributions:

$$p_G = \prod_{i=1}^{g} p_{\mathscr{C}_i} \left(\prod_{j=2}^{g} p_{\mathscr{S}_j} \right)^{-1} = p_{\mathscr{C}_1} \prod_{i=2}^{g} p_{\mathscr{C}_i|\mathscr{S}_i}, \tag{3}$$

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 $\Theta = \left(P_{\mathscr{C}_1}, P_{\mathscr{C}_2|\mathscr{S}_2},\ldots,P_{\mathscr{C}_g|\mathscr{S}_g}\right)$. We define a prior distribution on Θ as follows: for each $P_{\mathscr{C}_i|\mathscr{S}_i}$ and for each value of $\mathscr{S}_i$ we set a Dirichlet distribution defined for each possible combination of values $\mathbf{y}_{\mathscr{C}} \in \{0,1\}^{|\mathscr{C}|}$ of the variables in $\mathscr{C}$. 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: $N \sim Poi(\lambda), \quad \lambda \sim Gamma(\alpha,\beta)$;
iii) Rissanen's prior ([9]), $\pi(N) \propto 2^{-\log^*(N)}$, where $\log^*(N)$ is the sum of the positive terms in the sequence $\{\log_2(N), \log_2(\log_2(N)), \ldots\}$.
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 uncatchable 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 $n_{x,y}$ 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:

$$\text{Model 1: } [XA][XB][XC][XD][XE], \quad \text{and Model 2: } [XABC][XD][XE].$$

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]*).

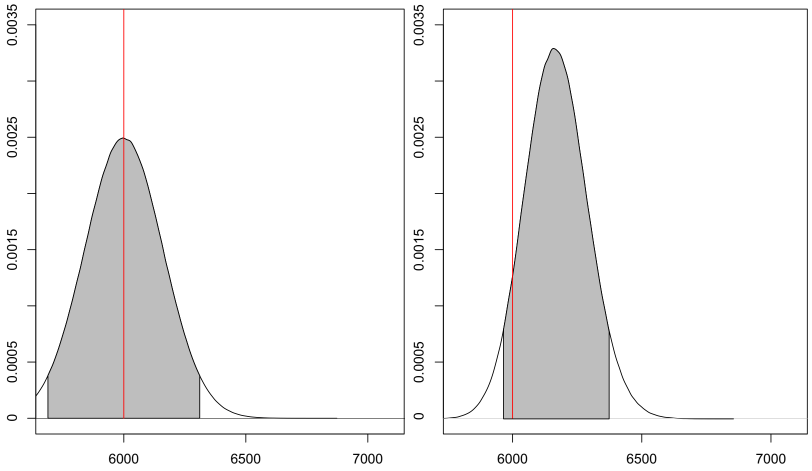

Fig. 1 Posterior distributions of N_1 under the generating model *[XABC][XD][XE]* (left) and under the all–second–order–interactions model (right). The orange line indicates the true value of N_1, the gray area the 95% HPD.

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.

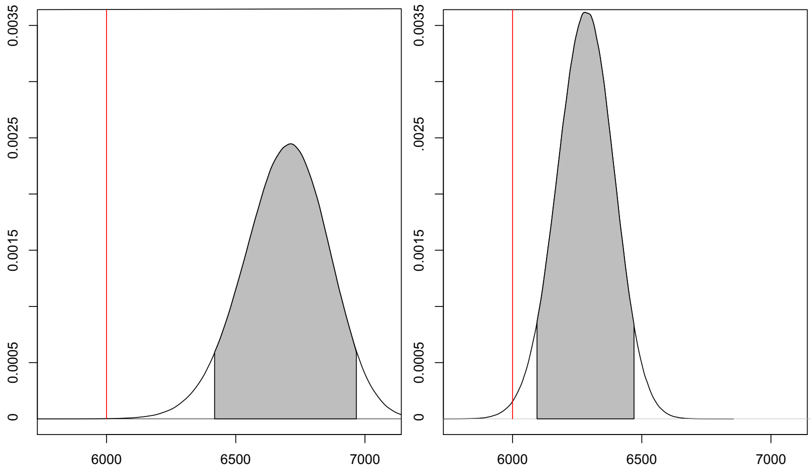

Fig. 2 Posterior distributions of N_1 under the CIA model, flat priors(left) and informative priors (right). The orange line indicates the true value of N_1, the gray area the 95% HPD.

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.