



UNIVERSITÀ DI PISA



**Sant'Anna**  
Scuola Universitaria Superiore Pisa



Consiglio Nazionale delle Ricerche

# Book of Short Papers

## SIS 2020



Società  
Italiana di  
Statistica

Editors: Alessio Pollice, Nicola Salvati and Francesco Schirripa Spagnolo

Copyright © 2020

PUBLISHED BY PEARSON

[WWW.PEARSON.COM](http://WWW.PEARSON.COM)

*ISBN 9788891910776*

# Contents

## Specialized sessions

Accounting for record linkage errors in inference (S2G-SIS).....	2
Probabilistic record linkage with less than three matching variables.	3
<i>Tiziana Tuoto and Marco Fortini</i>	
Advanced methods for measuring and communicating uncertainty in official statistics .....	9
A model for measuring the accuracy in spatial price statistics using scanner data.	10
<i>Ilaria Benedetti and Federico Crescenzi</i>	
Communication of Uncertainty of Official Statistics.	16
<i>Edwin de Jonge and Gian Luigi Mazzi</i>	
Measuring uncertainty for infra-annual macroeconomic statistics.	22
<i>George Kapetanios, Massimiliano Marcellino and Gian Luigi Mazzi</i>	
Bayesian methods in biostatistics .....	27
Network Estimation of Compositional Data.	28
<i>Nathan Osborne, Christine B. Peterson and Marina Vannucci</i>	
Using co-data to empower genomics-based prediction and variable selection.	34
<i>Magnus M. Münch, Mirrelijin M. van Nee and Mark A. van de Wiel</i>	
Data integration versus privacy protection: a methodological challenge? .....	40
Statistical Disclosure Control for Integrated Data.	41
<i>Natalie Shlomo</i>	
The Integrated System of Statistic Registers: first steps towards facing privacy issues.	47
<i>Mauro Bruno and Roberta Radini</i>	
Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics.	53
<i>Fabio Ricciato, Kostas Giannakouris, Albrecht Wirthmann and Martina Hahn</i>	
Designing adaptive clinical trials .....	59
Optimal designs for multi-arm exponential trials.	60
<i>Rosamarie Frieri and Marco Novelli</i>	
Education: students' mobility and labour market.....	66
From measurement to explanatory approaches: an assessment of the attractiveness of the curricula programs supplied by Italian universities.	67
<i>Isabella Sullis, Silvia Columbu and Mariano Porcu</i>	
Pull factors for university students' mobility: a gravity model approach.	73
<i>Giovanni Boscaïno and Vincenzo Giuseppe Genova</i>	
Spatial autoregressive gravity models to explain the university student mobility in Italy.	79
<i>Silvia Bacci, Bruno Bertaccini and Chiara Bocci</i>	

<b>Environmental Statistics (GRASPA-SIS) .....</b>	<b>85</b>
A Time Clustering Model for Spatio-Temporal Data. <i>Clara Grazian, Gianluca Mastrantonio and Enrico Bibbona</i>	86
Reconstruction of sparsely sampled functional time series using frequency domain functional principal components. <i>Amira Elayouty, Marian Scott and Claire Miller</i>	93
<b>Methods for High Dimensional Compositional Data Analysis .....</b>	<b>98</b>
Algorithms for compositional tensors of third-order. <i>Violetta Simonacci</i>	99
High-dimensional regression with compositional covariates: a robust perspective. <i>Gianna Serafina Monti and Peter Filzmoser</i>	105
Three-way compositional analysis of energy intensity in manufacturing. <i>Valentin Todorov and Violetta Simonacci</i>	111
<b>Modern Statistics for Physics Discoveries .....</b>	<b>117</b>
Identification of high-energy $\lambda$ -ray sources via nonparametric clustering. <i>Giovanna Menardi, Denise Costantin, and Federico Ferraccioli</i>	118
Statistical Analysis of Macroseismic Data for a better Evaluation of Earthquakes Attenuation Laws. <i>Marcello Chiodi, Antonino D'Alessandro, Giada Adelfio and Nicoletta D'Angelo</i>	124
<b>Network Modelling in Biostatistics.....</b>	<b>130</b>
Natural direct and indirect relative risk for mediation analysis. <i>Monia Lupparelli and Alessandra Mattei</i>	131
<b>New issues on multivariate and univariate quantile regression .....</b>	<b>137</b>
Mixtures of quantile regressions for longitudinal data: an R package. <i>Maria Francesca Marino, Maria Giovanna Ranalli and Marco Alfò</i>	138
Multivariate Mixed Hidden Markov Model for joint estimation of multiple quantiles. <i>Luca Merlo, Lea Petrella and Nikos Tzavidis</i>	144
<b>Recent methodological advances in finite mixture modeling with applications (CLADAG-SIS) .....</b>	<b>150</b>
Aggregating Gaussian mixture components. <i>Roberto Rocci</i>	151
Local and overall coefficients of determination for mixtures of generalized linear models. <i>Roberto Di Mari, Salvatore Ingrassia and Antonio Punzo</i>	157
<b>Statistical Analysis of Satellite Data (SDS-SIS) .....</b>	<b>163</b>
Functional Data Analysis for Interferometric Synthetic Aperture Radar Data Post-Processing: The case of Santa Barbara mud volcano. <i>Matteo Fontana, Alessandra Menafoglio, Francesca Cigna and Deodato Tapete</i>	164
Recent Contributions to the Understanding of the Uncertainty in Upper-Air Reference Measurements. <i>Alessandro Fassò</i>	170
<b>Statistical models and methods for Business and Industry .....</b>	<b>176</b>
Modelling and monitoring of complex 3D shapes: a novel approach for lattice structures. <i>Bianca Maria Colosimo, Marco Grasso and Federica Garghetti</i>	177
Open data powered territorial planning - Case study: The Turin historical center. <i>Silvia Casagrande, Gianmaria Origi, Alberto Pasanisi, Martina Tamburini, Pascal Terrien, Tania Cerquitelli and Alfonso Capozzoli</i>	183
Process optimization in Industry 4.0: Are all data analytics models useful? <i>Alberto Ferrer</i>	189

Technology and demographic behaviours (AISP-SIS) .....	195
Internet and the Timing of Births.	196
<i>Maria Sironi, Osea Giuntella and Francesco C. Billari</i>	
The Internetization of Marriage: Effects of the Diffusion of High-Speed Internet on Marriage, Divorce, and Assortative Mating.	202
<i>Francesco C. Billari, Osea Giuntella and Luca Stella</i>	

## Solicited Sessions

Advanced Statistical Methods in Health Analytics .....	209
Assessing the impact of the intermediate event in a non-markovian illness-death model.	210
<i>Davide Paolo Bernasconi, Elena Tassistro, Maria Grazia Valsecchi and Laura Antolini</i>	
Big data and AI: challenges and opportunities in healthcare.	216
<i>Vieri Emiliani, Gian Luca Cattani and Fabrizio Selmi</i>	
Statistical methodology for volume-outcome studies.	222
<i>Marta Fiocco and Floor van Oudenhoven</i>	
Advances in textual data mining .....	228
Distance measures for exploring pairs of novels in a large corpus of Italian literature.	229
<i>Matilde Trevisani and Arjuna Tuzzi</i>	
Supervised vs Unsupervised Latent Dirichlet Allocation: topic detection in lyrics.	235
<i>Mariangela Sciandra, Alessandro Albano and Irene Carola Spera</i>	
Advances in the interaction between artificial intelligence and official statistics .....	241
Automated Land Cover Maps from Satellite Imagery by Deep Learning.	242
<i>Fabrizio De Fausti, Francesco Pugliese and Diego Zardetto</i>	
CROWD4SDG: Crowdsourcing for sustainable developments goals.	248
<i>Barbara Pernici</i>	
Permanent Population Census: evaluation of the effects of regional strategies on the process efficiency. The direct experience of Tuscany.	253
<i>Linda Porciani, Luisa Francovich, Luca Faustini and Alessandro Valentini</i>	
Capture-recapture methods .....	259
Bayesian Model Averaging for Latent Class Models in Capture-Recapture.	260
<i>Davide Di Cecco</i>	
Combining "signs of life" and survey data through latent class models to consider over-coverage in Capture-Recapture estimates of population counts.	266
<i>Marco Fortini, Antonella Bernardini, Marco Caputi and Nicoletta Cibella</i>	
Population size estimation with interval censored counts and external information.	272
<i>Alessio Farcomeni</i>	
Changes in environment extremes and their impacts .....	278
FPCA Clustering of rainfall events.	279
<i>Gianluca Sottile, Antonio Francipane, Leonardo Noto and Giada Adelfio</i>	
Trends in rainfall extremes in the Venice lagoon catchment.	285
<i>Ilaria Prosdocimi and Carlo Gaetan</i>	

<b>Copulas: models and inference .....</b>	<b>291</b>
Analysis of district heating demand through different copula-based approaches. <i>F. Marta L. Di Lascio and Andrea Menapace</i>	292
CoVaR and backtesting: a comparison between a copula approach and parametric models. <i>Michele Leonardo Bianchi, Giovanni De Luca and Giorgia Riveccio</i>	298
Estimating Asymmetric Dependence via Empirical Checkerboard Copulas. <i>Wolfgang Trutschnig and Florian Griessenberger</i>	304
Strong Convergence of Multivariate Maxima. <i>Michael Falk, Simone A. Padoan and Stefano Rizzelli</i>	310
<b>Data Science: when different expertise meet .....</b>	<b>316</b>
Bayesian stochastic modelling of the temporal evolution of seismicity. <i>Elisa Varini and Renata Rotondi</i>	317
Cluster Analysis for the Characterization of Residential Personal Exposure to ELF Magnetic Field. <i>Gabriella Tognola, Silvia Gallucci, Marta Bonato, Emma Chiaramello, Isabelle Magne, Martine Souques, Serena Fiocchi, Marta Parazzini and Paolo Ravazzani</i>	323
Statistical Assessment and Validation of Ship Response in High Sea State by Computational Fluid Dynamics. <i>Andrea Serani, Matteo Diez and Frederick Stern</i>	328
Uncertainty Quantification for PDEs with random data using the Multi-Index Stochastic Collocation method. <i>Lorenzo Tamellini and Joakim Beck</i>	334
<b>Emerging challenges in official statistics: new data sources and methods .....</b>	<b>340</b>
Small area poverty indicators adjusted using local spatial price indices. <i>Stefano Marchetti, Luigi Biggeri, Caterina Giusti and Monica Pratesi</i>	341
Smart solutions for trusted smart statistics: the European big data hackathon experience. <i>Francesco Amato, Mauro Bruno, Tania Cappadozzi, Fabrizio De Fausti and Manuela Michelini</i>	347
The ESSnet Project Smart Surveys: new data sources and tools for Surveys of Official Statistics	353
<b>Factorial and dimensional reduction methods for the construction of indicators for evaluation (SVQS-SIS).....</b>	<b>359</b>
A comparison of MBC with CLV and PCovR methods for dimensional reduction of the soccer players' performance attributes. <i>Maurizio Carpita, Enrico Ciavolino and Paola Pasca</i>	360
A framework of cumulated chi-squared type statistics for ordered correspondence analysis. New tools and properties. <i>Antonello D'Ambra, Pietro Amenta and Luigi D'Ambra</i>	366
Exploring drug consumption via an ultrametric correlation matrix. <i>Giorgia Zaccaria and Maurizio Vichi</i>	372
Ranking extraction in ordinal multi-indicator systems. <i>Marco Fattore and Alberto Arcagni</i>	378
<b>Gender statistics .....</b>	<b>384</b>
Gender differences in Italian STEM degree courses: a discrete-time competing-risks model. <i>Marco Enea and Massimo Attanasio</i>	385
Some Challenges and Results in Measuring Gender Inequality. <i>Fabio Crescenzi and Francesco Di Pede</i>	391

<b>How Deep is Your Plot? Young SIS and deep statistical learning (ySIS)..</b>	<b>397</b>
A modal approach for clustering matrices.	398
<i>Federico Ferraccioli and Giovanna Menardi</i>	
A Note on Detection of Perturbations in Biological Networks.	404
<i>Vera Djordjilović</i>	
Bayesian inference for DAG-probit models.	410
<i>Federico Castelletti</i>	
Variational Bayes for Gaussian Factor Models under the Cumulative Shrinkage Process.	416
<i>Sirio Legramanti</i>	
<b>Measuring poverty and vulnerability .....</b>	<b>421</b>
Choosing the vulnerability threshold using the ROC curve.	422
<i>Chiara Gigliarano and Conchita D'Ambrosio</i>	
<b>New advances in applications, a Bayesian nonparametric perspective .....</b>	<b>428</b>
Bayesian Mixture Models for Latent Class Analysis.	429
<i>Raffaele Argiento, Bruno Bodin and Maria De Iorio</i>	
<b>Non-Parametric Inference and Forecasting of Functional and Object Data .....</b>	<b>435</b>
An interpretable estimator for the function-on-function linear regression model with application to the Canadian weather data.	436
<i>Fabio Centofanti and Matteo Fontana</i>	
Statistical process monitoring of multivariate profiles from ship operating conditions.	440
<i>Christian Capezza</i>	
<b>Prior choice in Bayesian Modelling (SISbayes) .....</b>	<b>446</b>
Bayesian Learning of Multiple Essential Graphs.	447
<i>Luca La Rocca, Federico Castelletti, Stefano Peluso, Francesco Claudio Stingo and Guido Consonni</i>	
Bayesian post-processing of Gibbs sampling output for variable selection.	453
<i>Stefano Cabras</i>	
Priors on precision parameters of IGRMF models.	459
<i>Aldo Gardini, Fedele Greco and Carlo Trivisano</i>	
<b>Sequence Analysis: methods and applications .....</b>	<b>465</b>
Internal migration, family formation and social stratification in Europe. A life course approach.	466
<i>Roberto Impicciatore, Gabriele Ballarino and Nazareno Panichella</i>	
<b>Socio economic integration of migrants .....</b>	<b>472</b>
A study on the characteristics of spouses who intermarry in Italy.	473
<i>Agnese Vitali and Romina Fraboni</i>	
<b>Statistical Analysis for mobility and transportation .....</b>	<b>479</b>
A multilevel Analysis of University attractiveness in the network flows from Bachelor to Master's degree.	480
<i>Silvia Columbu and Ilaria Primerano</i>	
Analysis of mobility data through a novel Cheng and Church algorithm for functional data.	486
<i>Marta Galvani, Agostino Torti and Alessandra Menafoglio</i>	
Bridge closures in a transportation network: analysis of the impacts in the region of Lombardy.	491
<i>Agostino Torti, Marika Arena, Giovanni Azzone, and Piercesare Secchi</i>	

<b>Statistical Methods and Applications in Social Network Analysis .....</b>	<b>496</b>
A clustering procedure for ego-networks data: an application to Italian elders living in couple. <i>Elvira Pelle and Roberta Pappadà</i>	497
Analysing the mediating role of a network: a Bayesian latent space approach. <i>Chiara Di Maria, Antonino Abbruzzo and Gianfranco Lovison</i>	503
Network-time autoregressive models for valued network panel. <i>Viviana Amati</i>	509
University student mobility flows and network data structures. <i>Maria Prosperina Vitale, Giuseppe Giordano and Giancarlo Ragozini</i>	515
<b>Statistical Methods in Psychometrics .....</b>	<b>521</b>
A simple probabilistic model to evaluate questionable interim analysis strategies. <i>Francesca Freuli and Luigi Lombardi</i>	522
Incorporating Expert Knowledge in Structural Equation Models: Applications in Psychological Research. <i>Gianmarco Altoè, Claudio Zandonella Callegher, Enrico Toffalini and Massimiliano Pastore</i>	528
Predicting social media addiction from Instagram profiles: A data mining approach. <i>Antonio Calcagni, Veronica Cortellazzo, Francesca Guizzo, Paolo Girardi, Natale Canale</i>	534
Structural entropy based modeling for psychological measurement. <i>Enrico Ciavolino, Mario Angelelli, Paola Pasca and Omar Carlo Gioacchino Gelo</i>	540
<b>Statistical modelling in environmental epidemiology .....</b>	<b>546</b>
A Time Varying Coefficient Model to Estimate the Short-Term Effects of Air Pollution on Human Health. <i>Pasquale Valentini, Luigi Ippoliti and Clara Grazian</i>	547
Joint Analysis of Short and Long-Term Effects of Air Pollution. <i>Annibale Biggeri, Dolores Catelan, Giorgia Stoppa and Corrado Lagazio</i>	551
<b>Statistical Modelling of Scientific Evidence for Forensic Investigation and Interpretation .....</b>	<b>557</b>
DNA mixtures with related contributors. <i>Peter J. Green and Julia Mortera</i>	558
Forensic Statistics: How to estimate life expectancy after injury. <i>Jane L Hutton</i>	564
The additional contribution of combining genetic evidence from multiple samples in a complex case. <i>Giampietro Lago</i>	570
The history of forensic inference and statistics: a thematic perspective. <i>Franco Taroni and Colin Aitken</i>	576
<b>Topological learning: interpretable representations of complex data.....</b>	<b>581</b>
Comparing Neural Networks via Generalized Persistence. <i>Mattia G. Bergomi and Pietro Vertechi</i>	582
On the topological complexity of decision boundaries. <i>António Leitão and Giovanni Petri</i>	588
Persistence-based Kernels for Data Classification. <i>Ulderico Fugacci</i>	594
Topological and Mixed-type learning of Brain Activity. <i>Tullia Padellini, Pierpaolo Brutti, Riccardo Giubilei</i>	600



## Contributed papers and Posters

<b>Bayesian Statistics</b> .....	<b>607</b>
A Bayesian approach for modelling dependence among mixture densities. <i>Mario Beraha, Matteo Pegoraro, Riccardo Peli and Alessandra Guglielmi</i>	608
A change of glasses strategy to solve the rare type match problem. <i>Giulia Cereda and Fabio Corradi</i>	614
A new prior distribution on the simplex: the extended flexible Dirichlet. <i>Roberto Ascari, Sonia Migliorati and Andrea Ongaro</i>	620
ABC model choice via mixture weight estimation. <i>Gianmarco Caruso, Luca Tardella and Christian P. Robert</i>	626
An ABC algorithm for random partitions arising from the Dirichlet process. <i>Mario Beraha and Riccardo Corradin</i>	632
Bayesian Inference of Undirected Graphical Models from Count Data. <i>Pier Giovanni Bissiri, Monica Chiogna and Nguyen Thi Kim Hue</i>	638
Bayesian IRT models in NIMBLE. <i>Sally Paganin, Chris Paciorek and Perry de Valpine</i>	644
Bayesian modelling of Facebook communities via latent factor models. <i>Emanuele Aliverti</i>	650
Bayesian nonparametric adaptive classification with robust prior information. <i>Francesco Denti, Andrea Cappozzo and Francesca Greselin</i>	655
Choosing the right tool for the job: a systematic analysis of general purpose MCMC software. <i>Mario Beraha, Giulia Gualtieri, Eugenia Villa, Riccardo Vitali and Alessandra Guglielmi</i>	661
Empirical Bayes estimation for mixture models. <i>Catia Scricciolo</i>	667
Improving ABC via Large Deviations Theory. <i>Cecilia Viscardi, Michele Boreale and Fabio Corradi</i>	673
Learning Bayesian Networks for Nonparanormal Data. <i>Flaminia Musella and Vincenzina Vitale</i>	679
Measuring well-being combining different data sources: a Bayesian networks approach. <i>Federica Cugnata, Silvia Salini and Elena Siletti</i>	685
Penalising the complexity of extensions of the Gaussian distribution. <i>Diego Battagliese and Brunero Liseo</i>	691
Predictive discrepancy of credible intervals for the parameter of the Rayleigh distribution. <i>Fulvio De Santis and Stefania Gubbiotti</i>	697
Small-area statistical estimation of claim risk. <i>Francesca Fortunato, Fedele Greco and Pierpaolo Cristaudo</i>	702
Subject-specific Bayesian Hierarchical model for compositional data analysis. <i>Matteo Pedone and Francesco C. Stingo</i>	708
Wasserstein consensus for Bayesian sample size determination. <i>Michele Cianfriglia, Tullia Padellini and Pierpaolo Brutti</i>	714
<b>Biostatistics</b> .....	<b>720</b>
A comparison of the CAR and DAGAR spatial random effects models with an application to diabetes rate estimation in Belgium. <i>Vittoria La Serra, Christel Faes, Niel Hens and Pierpaolo Brutti</i>	721
A functional approach to study the relationship between dynamic covariates and survival outcomes: an application to a randomized clinical trial on osteosarcoma. <i>Marta Spreafico, Francesca Ieva and Marta Fiocco</i>	727

A Statistical Approach to the Alignment of fMRI Data. <i>Angela Andreella, Ma Feilong, Yaroslav Halchenko, James Haxby and Livio Finos</i>	733
Adaptive clinical trials: Bayesian decision-theoretic and frequentist approaches for cost-effectiveness analysis. <i>Martin Forster and Marco Novelli</i>	739
Bootstrap corrected Propensity Score: Application for Anticoagulant Therapy in Haemodialysis Patients. <i>Maeregu W. Arisido, Fulvia Mecatti and Paola Rebora</i>	745
Combining multiple sources to overcome misclassification bias in epidemiological database studies. <i>Francesca Beraldi, Rosa Gini, Emanuela Dreassi, Leonardo Grilli and Carla Rampichini</i>	751
Deep Sparse Autoencoder-based Feature Selection for SNPs Validation in Prostate Cancer Radiogenomics. <i>Michela Carlotta Massi, Francesca Ieva, Anna Maria Paganoni, Andrea Manzoni, Paolo Zunino, Nicola Rares Franco, Tiziana Rancati and Catharine West</i>	756
Graphical models for count data: an application to single-cell RNA sequencing. <i>Nguyen Thi Kim Hue, Monica Chiogna and Davide Rizzo</i>	762
Interregional mobility, socio-economic inequality and mortality among cancer patients. <i>Claudio Rubino, Mauro Ferrante, Antonino Abbruzzo, Giovanna Fantaci and Salvatore Scondotto</i>	768
PET radiomics-based lesions representation in Hodgkin lymphoma patients. <i>Lara Cavinato, Martina Sollini, Margarita Kirienko, Matteo Biroli, Francesca Ricci, Letizia Calderoni, Elena Tabacchi, Cristina Nanni, Pier Luigi Zinzani, Stefano Fanti, Anna Guidetti, Alessandra Alessi, Paolo Corradini, Ettore Seregni, Carmelo Carlo-Stella, Arturo Chiti and Francesca Ieva</i>	774
Prediction of late radiotherapy toxicity in prostate cancer patients via joint analysis of SNPs sequences. <i>Nicola Rares Franco, Michela Carlotta Massi, Francesca Ieva, Anna Maria Paganoni, Andrea Manzoni, Paolo Zunino, Tiziana Rancati and Catharine West</i>	780
Predictive versus posterior probabilities for phase II trial monitoring. <i>Valeria Sambucini</i>	785
Profile networks for precision medicine. <i>Andrea Lazerini, Monia Lupporelli and Francesco C. Stingo</i>	791
Proton-Pump Inhibitor Provider Profiling via Funnel Plots and Poisson Regression. <i>Dario Delle Vedove, Francesca Ieva and Anna Maria Paganoni</i>	797
Selecting optimal thresholds in ROC analysis with clustered data. <i>Duc Khanh To, Gianfranco Adimari and Monica Chiogna</i>	803
<b>Environment, Physics and Engineering .....</b>	<b>809</b>
A hidden semi-Markov model for segmenting environmental toroidal data. <i>Francesco Lagona and Antonello Maruotti</i>	810
An experimental analysis on quality and security about green communication. <i>Vito Santarcangelo, Emilio Massa, Davide Scintu, Michele Di Lecce and Massimiliano Giacalone</i>	816
An improved sensitivity-data based method for probabilistic ecological risk assessment. <i>Sonia Migliorati and Gianna Serafina Monti</i>	822
Comparing predictive distributions in EMOS. <i>Giummolè Federica and Mameli Valentina</i>	828
Compositional analysis of fish communities in a fast changing marine ecosystem. <i>Pierfrancesco Alaimo Di Loro, Marco Mingione, Giovanna Jona Lasinio, Sara Martino and Francesco Colloca</i>	834
FDA dimension reduction techniques and components separation in Fourier-transform infrared spectroscopy. <i>Francesca Di Salvo, Elena Piacenza and Delia Francesca Chillura Martino</i>	840
Functional Data Analysis for Spectroscopy Data. <i>Mara S. Bernardi, Matteo Fontana, Alessandra Menafoglio, Diego Perugini, Alessandro Pisello, Marco Ferrari, Simone De Angelis, Maria Cristina De Sanctis and Simone Vantini</i>	846
Functional graphical model for spectrometric data analysis. <i>Laura Codazzi, Alessandro Colombi, Matteo Gianella, Raffaele Argiento, Lucia Paci and Alessia Pini</i>	852
Local LGCP estimation for spatial seismic processes. <i>Nicoletta D'Angelo, Marianna Siino, Antonino D'Alessandro and Giada Adelfio</i>	857

Observation-driven models for storm counts. <i>Mirko Armillotta, Alessandra Luati and Monia Lupparelli</i>	863
Statistical control of complex geometries, with application to Additive Manufacturing. <i>Riccardo Scimone, Tommaso Taormina, Bianca Maria Colosimo, Marco Grasso, Alessandra Menafoglio, Piercesare Secchi</i>	869
Tree attributes map by 3P sampling in a design-based framework. <i>Lorenzo Fattorini and Sara Franceschi</i>	875
Unsupervised classification of texture images by gray-level spatial dependence matrices and genetic algorithms. <i>Roberto Baragona and Laura Bocci</i>	880
<b>Finance, business and official statistics .....</b>	<b>886</b>
A discrete choice approach to analyze contractual attributes in the durum wheat sector in Italy. <i>Stefano Ciliberti, Simone Del Sarto, Giulia Pastorelli, Angelo Frascarelli and Gaetano Martino</i>	887
A fuzzy approach to the measurement of the employment rate. <i>Bruno Cheli, Alessandra Coli and Andrea Regoli</i>	893
A proposal to model credit risk contagion using network count-based models. <i>Arianna Agosto and Daniel Felix Ahelegbey</i>	898
A similarity matrix approach to empower ESCO interfaces for testing, debugging and in support of users' experience. <i>Adham Kahlawi, Cristina Martelli, Lucia Buzzigoli, Laura Grassini</i>	904
Adding MIDAS terms to Linear ARCH models in a Quantile Regression framework. <i>Vincenzo Candila and Lea Petrella</i>	910
Company requirements in Italian tourism sector: an analysis for profiles. <i>Paolo Mariani, Andrea Marletta, Lucio Masserini and Mariangela Zenga</i>	916
Determinants of Firms' Default Risk after the 2008 and 2011 Economic Crises: a Latent Growth Models Approach. <i>Lucio Masserini, Matilde Bini and Alessandro Zeli</i>	921
Double Asymmetric GARCH-MIDAS model - new insights and results. <i>Alessandra Amendola, Vincenzo Candila and Giampiero M. Gallo</i>	927
European SMEs and Circular Economy Activities: Evaluating the Advantage on Firm Performance through the Estimation of Average Treatment Effects. <i>Luca Secondi</i>	933
Financial Spillover Measures to Assess the Stability of Basket-based Stablecoins. <i>Paolo Pagnottoni</i>	939
Forecasting Banknote Flows in Bdl Branches: Speed-up with Machine Learning. <i>Marco Brandi, Monica Fusaro, Tiziana Laureti and Giorgia Rocco</i>	945
Fully reconciled GDP forecasts from Income and Expenditure sides. <i>Luisa Bisaglia, Tommaso Di Fonzo and Daniele Girolimetto</i>	951
GLASSO Estimation of Commodity Risks. <i>Beatrice Foroni, Saverio Mazza, Giacomo Morelli and Lea Petrella</i>	957
Measuring the Effect of Unconventional Policies on Stock Market Volatility. <i>Giampiero M. Gallo, Demetrio Lacava and Edoardo Otranto</i>	963
Multidimensional versus unidimensional poverty measurement. <i>Michele Costa</i>	969
Multiple outcome analysis of European Agriculture in 2000-2016: a latent class multivariate trajectory approach. <i>Alessandro Magrini</i>	975
Nowcasting GDP using mixed-frequency based composite confidence indicators. <i>Maria Carannante, Raffaele Mattered, Michelangelo Misuraca, Germana Scepi and Maria Spano</i>	981
On the tangible and intangible assets of Initial Coin Offerings. <i>Paola Cerchiello and Anca Mirela Toma</i>	987

Seasonality variation of electricity demand: decompositions and tests. <i>Luigi Grossi and Mauro Mussini</i>	993
SMEs circular economy practices in the European Union: Implications for sustainability. <i>Nunzio Tritto, José G. Dias and Francesca Bassi</i>	999
Tax Incentives' Effect on the Provision of Occupational Welfare in Italian Enterprises. <i>Alessandra Righi</i>	1005
The determinants of eco-innovation: a country comparison using the community innovation survey. <i>Ida D'Attoma and Silvia Pacei</i>	1011
World ranking of urban sustainability through composite indicators. <i>Elena Grimaccia, Alessia Naccarato and Silvia Terzi</i>	1017
<b>Machine Learning and Data Science.....</b>	<b>1023</b>
A novel approach for Artificial Intelligence through Lorenz zonoids and Shapley Values. <i>Paolo Giudici and Emanuela Raffinetti</i>	1024
A warning signal for variable importance interpretation in tree-based algorithms. <i>Anna Gottard and Giulia Vannucci</i>	1030
Assessment of the effectiveness of digital flyers: analysis of viewing behavior using eye tracking. <i>Gianpaolo Zammarchi, Claudio Conversano and Francesco Mola</i>	1036
At risk mental status analysis: a comparison of model selection methods for ordinal target variable. <i>Elena Ballante, Silvia Molteni, Martina Mensi and Silvia Figini</i>	1042
Categorical Encoding for Machine Learning. <i>Agostino Di Ciaccio</i>	1048
Dynamic Quantile Regression Forest. <i>Mila Andreani and Lea Petrella</i>	1054
Estimating the UK Sentiment Using Twitter. <i>Stephan Schlosser, Daniele Toninelli and Michela Cameletti</i>	1059
Forecasting local rice prices from crowdsourced data in Nigeria. <i>Ilaria Lucrezia Amerise and Gloria Solano Hermosilla</i>	1065
Generalized Mixed Effects Random Forest: does Machine Learning help in predicting university student dropout? <i>Massimo Pellagatti, Chiara Masci, Francesca Ieva and Anna Maria Paganoni</i>	1071
HateViz: a textual dashboard Twitter data-driven. <i>Emma Zavarrone, Maria Gabriella Grassia, Marina Marino, Rocco Mazza and Nicola Canestrari</i>	1077
How to perform cyber risk assessment via cumulative logit models. <i>Silvia Facchinetti, Silvia Angela Osmetti and Claudia Tarantola</i>	1083
Machine learning prediction for accounting system. <i>Chiara Bardelli and Silvia Figini</i>	1087
Teaching statistics: an assessment framework based on Multidimensional IRT and Knowledge Space Theory. <i>Cristina Davino, Rosa Fabbriatore, Carla Galluccio, Daniela Pacella, Domenico Vistocco, Francesco Palumbo</i>	1093
The weight of words: textual data versus sentiment analysis in stock returns prediction. <i>Riccardo Ferretti and Andrea Sciandra</i>	1099
Unsupervised Energy Trees: clustering with complex and mixed-type variables. <i>Riccardo Giubilei, Tullia Padellini and Pierpaolo Brutti</i>	1105
Using anchoring vignettes to adjust self-reported life satisfaction: a nonparametric approach leading to a Semantic Differential scale. <i>Sara Garbin, Serena Berretta, Maria Iannario and Omar Paccagnella</i>	1111
Variable selection for robust model-based learning from contaminated data. <i>Andrea Cappozzo, Francesca Greselin and Thomas Brendan Murphy</i>	1117

Variable Selection in Text Regressions: Back to Lasso? <i>Marzia Freo and Alessandra Luati</i>	1123
Web Usage Mining and Website Effectiveness. <i>Maria Francesca Cracolici and Furio Urso</i>	1129
<b>Models and methods - Categorical, Ordinal, Rank Data .....</b>	<b>1135</b>
Aberration for the analysis of two-way contingency tables. <i>Roberto Fontana and Fabio Rapallo</i>	1136
An investigation of the paradoxical behaviour of $\kappa$ -type inter-rater agreement coefficients for nominal data. <i>Amalia Vanacore and Maria Sole Pellegrino</i>	1142
Analyzing faking-good response data: Combination of a Replacement and a Binomial (CRB) distribution approach. <i>Luigi Lombardi and Antonio Calcagni</i>	1148
BOD – min range: A Robustness Analysis Method for Composite Indicators. <i>Emiliano Seri, Leonardo Salvatore Alaimo and Vittoria Carolina Malpassuti</i>	1154
Comparing classifiers for ordinal variables. <i>Silvia Golia and Maurizio Carpita</i>	1160
Discovering Interaction Effects Between Subject-Specific Covariates: A New Probabilistic Approach For Preference Data. <i>Alessio Baldassarre, Claudio Conversano, Antonio D'Ambrosio, Mark De Rooij and Elise Dusseldorp</i>	1166
Hybrid random forests for ordinal data. <i>Rosaria Simone and Gerhard Tutz</i>	1171
Model-based approach to biclustering ordinal data. <i>Monia Ranalli and Francesca Martella</i>	1177
New algorithms and goodness-of-fit diagnostics for ranked data modelling with the Extended Plackett-Luce distribution. <i>Cristina Mollica and Luca Tardella</i>	1183
Non-metric unfolding on augmented data matrix: a copula-based approach. <i>Marta Nai Ruscone and Antonio D'Ambrosio</i>	1189
Ordinal probability effect measures for dyadic analysis in cumulative models. <i>Maria Iannario and Domenico Vistocco</i>	1194
Simulated annealing for maximum rater agreement. <i>Fabio Rapallo and Maria Piera Rogantin</i>	1200
<b>Models and methods – Regression.....</b>	<b>1206</b>
A Clusterwise regression method for Distributional-valued Data. <i>Rosanna Verde, Francisco de A. T. de Carvalho and Antonio Balzanella</i>	1207
A nonparametric approach for nonlinear variable screening in high-dimensions. <i>Francesco Giordano, Sara Milito and Lucia Maria Parrella</i>	1213
Adjusted scores for inference in negative binomial regression. <i>Euloge C. Kenne Pagui, Alessandra Salvan and Nicola Sartori</i>	1219
Estimation of the treatment effect variance in a difference-in-differences framework. <i>Marco Doretti and Giorgio E. Montanari</i>	1224
Exploring multicollinearity in quantile regression. <i>Cristina Davino, Tormod Naes, Rosaria Romano and Domenico Vistocco</i>	1230
Generalized M-quantile random effects model. <i>Francesco Schirripa Spagnolo and Vincenzo Mauro</i>	1236
Goodness-of-fit assessment in linear quantile regression. <i>Ilaria Lucrezia Amerise and Agostino Tarsitano</i>	1242
Joint Redundancy Analysis by a multivariate linear predictor. <i>Laura Marcis and Renato Salvatore</i>	1248

M-quantile regression shrinkage and selection via the lasso. <i>M. Giovanna Ranalli, Lea Petrella and Francesco Pantalone</i>	1254
New insights into the Conditioning and Gain Score approaches in multilevel analysis. <i>Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto and Carla Rampichini</i>	1260
Simultaneous confidence regions and curvature measures in nonlinear models. <i>Claudia Furlan and Cinzia Mortarino</i>	1265
<b>Models and methods – Sampling .....</b>	<b>1271</b>
Design-based consistency of the Horvitz-Thompson estimator for spatial populations. <i>Lorenzo Fattorini, Marzia Marcheselli, Caterina Pisani and Luca Pratelli</i>	1272
Empirical likelihood in the statistical matching for informative samples. <i>Daniela Marella and Danny Pfeffermann</i>	1278
Evaluating a Hybrid One-Stage Snowball Sampling through Bootstrap Method on a Simulated Population. <i>Venera Tomaselli and Giulio Giacomo Cantone</i>	1284
How optimal subsampling depends on guessed parameter values. <i>Laura Deldossi and Chiara Tommasi</i>	1290
Indicators for risk of selection bias in non-probability samples. <i>Emilia Rocco and Alessandra Petrucci</i>	1296
On the behaviour of the maximum likelihood estimator for exponential models under a fixed and a two-stage design. <i>Caterina May and Chiara Tommasi</i>	1302
Pseudo-population based resamplings for two-stage design. <i>Pier Luigi Conti, Daniela Marella and Vincenzina Vitale</i>	1308
<b>Models and methods - Theoretical Issues in Statistical Inference .....</b>	<b>1314</b>
A new mixture model for three-way data. <i>Salvatore D. Tomarchio, Antonio Punzo and Luca Bagnato</i>	1315
A Sequential Test for the C <sub>pmk</sub> Index. <i>Michele Scagliarini</i>	1320
Probability Interpretations and the Selection of the Most Effective Statistics Method. <i>Paolo Rocchi</i>	1326
Robust Composite Inference. <i>Valentina Mamei, Monica Musio, Erlis Ruli and Laura Ventura</i>	1332
Statistical hypothesis testing within the Generalized Error Distribution: Comparing the behavior of some nonparametric techniques. <i>Massimiliano Giacalone and Demetrio Panarello</i>	1338
Stochastic dependence with discrete copulas. <i>Fabrizio Durante and Elisa Perrone</i>	1344
<b>Models and methods - Time Series and Longitudinal Data.....</b>	<b>1350</b>
Bootstrap test in Poisson-INAR models. <i>Lucio Palazzo and Riccardo Ievoli</i>	1351
Continuous Time-Interaction Processes for Population Size Estimation. <i>Linda Altieri, Alessio Farcomeni, Danilo Alunni Fegatelli and Francesco Palini</i>	1357
Longitudinal data analysis using PLS-PM approach. <i>Rosanna Cataldo, Corrado Crocetta, Maria Gabriella Grassia and Marina Marino</i>	1363
Long-memory models for count time series. <i>Luisa Bisaglia, Massimiliano Caporin and Matteo Grigoletto</i>	1369

Combining multiple frequencies in Realized GARCH models. <i>Antonio Naimoli and Giuseppe Storti</i>	1375
Models with Time-Varying Parameters for Realized Covariance. <i>Luc Bauwens and Edoardo Otranto</i>	1381
Pitman-Yor mixture models for survival data stratification. <i>Riccardo Corradin, Luis Enrique Nieto Barajas and Bernardo Nipoti</i>	1387
Prediction is not everything, but everything is prediction. <i>Leonardo Egidi</i>	1393
The Generalized Dynamic Mixtures of Factor Analyzers for clustering multivariate longitudinal data. <i>Francesca Martella, Antonello Maruotti and Francesco Tursini</i>	1399
Trends and long-run relations in cointegrated time series observed with noise. <i>Angelica Gianfreda, Paolo Maranzano, Lucia Parisio and Matteo Pelagatti</i>	1405
<b>Population and society .....</b>	<b>1411</b>
A dimensionality assessment of refugees' vulnerability through an Item Response Theory approach. <i>Simone Del Sarto, Michela Gnalzi, Yara Maasri and Edouard Legoupil</i>	1412
Accounting for Interdependent Risks in Vulnerability Assessment of Refugees. <i>Daria Mendola, Anna Maria Parroco and Paolo Li Donni</i>	1418
Active ageing in China: What are the domains that most affect life satisfaction in the elderly? <i>Ilaria Rocco</i>	1424
Analyzing the waiting time of academic publications: a survival model. <i>Francesca De Battisti, Giuseppe Gerardi, Giancarlo Manzi and Francesco Porro</i>	1430
Clustering of food choices in a large sample of students using university canteen. <i>Valentina Lorenzoni, Isotta Triulzi, Irene Martinucci, Letizia Toncelli, Michela Natilli and Roberto Barale, Giuseppe Turchetti</i>	1436
Cruise passengers' expenditure at destinations: Review of survey techniques and data collection. <i>Caterina Sciortino, Stefano De Cantis, Mauro Ferrante and Szilvia Gyimóthy</i>	1442
Educational integration of foreign citizen children in Italy: a synthetic indicator. <i>Alessio Buonomo, Stefania Capecchi and Rosaria Simone</i>	1448
Estimating the Change in Housework Time of the Italian Woman after the Retirement of the Male Partner: An Approach Based on a Two-Regime Model Estimated by ML. <i>Giorgio Calzolari, Maria Gabriella Campolo, Antonino Di Pino and Laura Magazzini</i>	1454
First and Second Year Careers of STEM Students in Italy: A Geographical Perspective. <i>Antonella D'Agostino, Giulio Ghellini and Gabriele Lombardi</i>	1460
Future Scenarios and Support Interventions for the Family: Involving Experts' Participation through a Mixed-Method Research Study. <i>Mario Bolzan, Simone Di Zio, Manuela Scioni and Morena Tartari</i>	1466
Gender and Monetary Policy Preferences: a Diff-in-Diff Approach. <i>Donata Favaro, Anna Giraldo and Ina Gollikja</i>	1472
Headcount based indicators and functions to evaluate the effectiveness of Italian university education. <i>Silvia Terzi and Francesca Petrarca</i>	1478
Identify the speech code through statistics: a data-driven approach. <i>Andrea Briglia, Massimo Mucciardi and Jérémi Sauvage</i>	1484
Inspecting cause-specific mortality curves by simplicial functional data analysis. <i>Marco Stefanucci and Stefano Mazzucco</i>	1490
Intertemporal decision making and childless couples. <i>Daniela Bellani, Bruno Arpino and Daniele Vignoli</i>	1495
Italian Households' Material Deprivation: Multi-Objective Genetic Algorithm approach for categorical variables. <i>Laura Bocci and Isabella Mingo</i>	1501



LI-CoD Model. From Lifespan Inequality to Causes of Death. <i>Andrea Nigri and Susanna Levantesi</i>	1507
Modeling Well-Being through PLS-SEM and K-M. <i>Venera Tomaselli, Mario Fordellone and Maurizio Vichi</i>	1513
News life-cycle: a multiblock approach to the study of information. <i>Rosanna Cataldo, Marco Del Mastro, Maria Gabriella Grassia, Marina Marino and Rocco Mazza</i>	1519
Short-term rentals in a tourist town. <i>Silvia Bacci, Bruno Bertaccini, Gianni Dugheri, Paolo Galli, Antonio Giusti and Veronica Sula</i>	1525
Sportstat: a playful activity to developing statistical literacy. <i>Alessandro Valentini and Francesca Paradisi</i>	1531
Statistical modeling for some features of Airbnb activity. <i>Giulia Contu and Luca Frigau</i>	1537
Tertiary students with migrant background: evidence from a cohort enrolled at Sapienza University. <i>Cristina Giudici, Donatella Vicar and Eleonora Trappolini</i>	1543
The Causal Effect of Immigration Policies on Income Inequality. <i>Irene Crimaldi, Laura Forastiere, Fabrizia Mealli and Costanza Tortù</i>	1549
The job condition of academic graduates: a joint longitudinal analysis of AlmaLaurea and Mandatory Notices of the Ministry of Labour. <i>Maria Veronica Dorgali, Silvia Bacci, Bruno Bertaccini and Alessandra Petrucci</i>	1557
The joint effect of childcare services and flexible female employment on fertility rate in Europe. <i>Viviana Cocuccio and Massimo Mucciardi</i>	1565
The Left Behind Generation: How the current Early School Leavers affect tomorrow's NEETs? <i>Giovanni De Luca, Paolo Mazzocchi, Claudio Quintano and Antonella Rocca</i>	1571
The probability to be employed of young adults of foreign origin. <i>Alessio Buonomo, Francesca Di Iorio and Salvatore Strozza</i>	1577
The risk of inappropriateness in geriatric wards: a comparison among the Italian regions. <i>Paolo Mariani, Andrea Marietta, Marcella Mazzoleni and Mariangela Zenga</i>	1583
The role of the accumulation of poverty and unemployment for health disadvantages. <i>Annalisa Busetta, Daria Mendola, Emanuela Struffolino and Zachary Van Winkle</i>	1589
Unemployment and fertility in Italy. A regional level data panel analysis. <i>Gabriele Ruiu and Marco Breschi</i>	1595
University drop out and mobility in Italy. First evidences on first level degrees. <i>Nicola Tedesco and Luisa Salaris</i>	1601
Worthiness-based Scale Quantifying. <i>Giulio D'Epifanio</i>	1607
Young people in Southern Italy and the phenomenon of immigration: what is their perception? <i>Nunziata Ribecco, Angela Maria D'Ugento and Angela Labarile</i>	1613





# Preface

The COVID-19 pandemic is putting our society under incredible health, emotional, and economic stress. Facing its harmful effects and their uncertainty, the Executive Board of the Italian Statistical Society (SIS) and the Local Organizing Committee, to ensure the highest level of safety for members and delegates, deliberated to cancel the 50th Meeting of the Italian Statistical Society originally planned to be held in Pisa in June 2020 and to postpone the conference to June 2021. The Executive Board and the Local Organizing Committee continue to monitor closely the pandemic evolving situation, and keep the members of SIS and the researchers informed about the potential new dates for the next meeting. To give value to the work of those who prepared their presentation for the conference, the Program Committee decided to publish the volume *Book of short papers - SIS 2020* despite the conference cancellation.

The conference program included 4 plenary sessions, 16 specialized sessions, 24 solicited sessions, 32 contributed sessions and the poster exhibition. Plenary sessions concerned with robust statistics, human longevity, statistical models for climate changes and small area estimation for educational poverty. The meeting had to host also 2 round tables on data privacy and innovation in statistics. Activities focused on topics of interest for a wider audience included two round tables on Teaching Statistics and on the SIS journal Statistical Methods & Applications, and the Stats Under the Stars (SUS6) competition for young statisticians. The SUS6 event attracted many sponsors from statistical, financial and editorial firms as well as numerous students. The conference committee had registered 345 accepted submissions, including 143 to be presented in invited plenary, specialized and solicited sessions, and 202 spontaneously submitted for oral and poster sessions.

This book includes most of the scientific contributions that had to be presented at the 50th Meeting of the Italian Statistical Society. It is organized into 49 chapters corresponding to 15 specialized, 23 solicited sessions, and to 11 general topics for contributed papers and posters. All 268 contributions provide a wide overview of the state-of-the-art of the subjects, from methodological and theoretical contributions, to applied works and case studies. The result is a very lively picture of the Italian statistical community with its international connections.

We would like to thank all contributors for having submitted their work to the conference, the members of the Program Committee and the extra reviewers for their efforts in this difficult period. Although the Conference did not take place, the organization went on until cancellation was decided for safety reasons. It would have been impossible without the joint effort of Università di Pisa, Scuola Superiore Sant'Anna and National Research Council of Pisa. Members these three institutions took part actively in the Local Organizing Committee. Finally we wish to express our gratitude to the publisher Pearson Italia for all the support received.

This book is our contribution to encourage the scientific community and the network of the Italian Statistical Society to go on and transform this difficult period into an opportunity of scientific debate for better statistics in a better world.

Alessio Pollice  
Università degli Studi di Bari Aldo Moro  
Chair of the Program Committee

Nicola Salvati  
Università di Pisa  
Chair of the Local Organizing Committee

Francesco Schirripa Spagnolo  
Università di Pisa

**Program Committee:** Alessio Pollice (Chair), Serena Arima, Marilena Barbieri, Alessandra Brazzale, Eugenio Brentari, Alessia Caponera, Antonio Lepore, Antonella Plaia, Tommaso Proietti, Marco Riani, Nicola Salvati, Pasquale Sarnacchiaro, Mauro Scanu, Manuela Stranges, Valentina Tocchioni, Simone Vantini, Massimo Ventrucci, Paola Vicard, Donatella Vicari.

**Local Organizing Committee:** Nicola Salvati (Chair), Gaia Bertarelli, Bruno Cheli, Alessandra Coli, Paolo Frumento, Fosca Giannotti, Caterina Giusti, Piero Manfredi, Stefano Marchetti, Lucio Masserini, Vincenzo Mauro, Barbara Pacini, Dino Pedreschi, Francesco Schirripa Spagnolo, Chiara Seghieri.

**Organizers of Specialized and Solicited Sessions:** Giada Adelfio, Bruno Arpino, Emanuele Aliverti, Nicoletta Balbo, Mara Bernardi, Silvia Bozza, Pierpaolo Brutti, Annalisa Busetta, Michela Cameletti, Carlo Cavicchia, Fabrizio Durante, Leonardo Egidi, Pietro D. Falorsi, Francesco Finazzi, Livio Finos, Stefania Galimberti, Michele Gallo, Caterina Giusti, Francesca Greselin, Alessandra Guglielmi, Francesca Ieva, Tiziana Laureti, Achille Lemmi, Brunero Liseo, Fabio Massimo Lo Verde, Daria Mendola, Roberta Pappadà, Lea Petrella, Alessandra Petrucci, Alessia Pini, Sabrina Prati, Maria Giovanna Ranalli, Davide Risso, Fabrizio Ruggeri, Silvana Salvini, Monica Scannapieco, Francesco Stingo, Luca Tardella, Grazia Vicario, Susanna Zaccarin, Maroussa Zagoraïou.

# Bayesian Model Averaging for Latent Class Models in Capture–Recapture

## *Model Averaging Bayesiano per Modelli a Classi Latenti in Cattura-Ricattura*

Davide Di Cecco

**Abstract** Model selection appears to be crucial in capture-recapture problems as it is common that different models with an equally good level of adaptation to the observed data lead to rather different estimates of the undercounts. We consider log–linear Latent Class Models as our capture-recapture model and propose Bayesian model averaging to overcome the difficulties of model selection within this class. We show that, by focusing on graphical decomposable models, we can design a simple Gibbs–based MCMC to sample over the space of eligible models.

**Abstract** *In problemi di cattura–ricattura, la selezione del modello risulta essere un aspetto delicato e difficoltoso. Non è inusuale, infatti, trovare modelli con valori di bontà di adattamento molto simili tra di loro che conducono a delle stime del numero di unità non catturate molto diverse. In questo lavoro trattiamo una famiglia estesa di modelli a classi latenti che rilassa l'ipotesi di indipendenza condizionata modellando le interazioni tra le variabili attraverso un modello log–lineare con una variabile latente. Per superare la difficoltà di scelta del modello all'interno di questa classe ampliata, proponiamo un model averaging Bayesiano. Mostriamo come, se ci limitiamo a considerare i modelli log-lineari decomponibili, è possibile costruire un semplice Gibbs sampler per ottenere la distribuzione a posteriori della numerosità della popolazione d'interesse.*

**Key words:** Bayesian Model Averaging, Latent Class Models, Capture-Recapture

## 1 Introduction

As pointed out by many authors, see, e.g., [4], the problem of estimating the size of a population in a capture-recapture model is essentially a problem of forecasting. As a consequence, it is not unusual that different models with a comparable level of

---

Davide Di Cecco

Sapienza University, viale del Castro Laurenziano 9, e-mail: davide.dicecco@uniroma1.it

goodness of fit lead to rather different estimates of the total population count. Given the lack of specific criteria for model choice, and the impossibility to validate the estimates, it is not unusual in capture-recapture practice to simply rule out a model resulting in unrealistic estimates. We think that a reliable procedure to deal with model selection in a more automatic way would certainly be of interest.

We treat the case of Multiple Record System, that is, the data consists of a set of capturing lists, usually originating from different sources, reporting partial listing of the same target population. In this setting it is common to assume different capture probabilities for the various sources. As a consequence, log-linear models are the tool of choice in capture-recapture modeling, and Latent Class Models (LCM) represent the natural extension when one wants to include unobserved heterogeneity. The use of LCM in capture-recapture dates back at least to [1] with many developments thereafter. The simplest formulation of LCM envisages the conditional independence assumption (CIA) which appears to be too restrictive in many situations. There are many proposals in literature to relax the CIA resulting in more flexible models. We focus on log-linear LCM where the additional dependencies are directly modeled by interaction parameters. Previous works on this class include [3], [13], [12]. We propose a Bayesian approach to the class as previously introduced in [6] and [7]. To overcome the difficulty of model selection within this class, we propose Bayesian model averaging to analyze the posterior distribution of the population count over a set of eligible models. Usually a full Bayesian approach to model averaging requires the use of a Reversible Jump algorithm ([8]) which is in general hard to implement. See [11] for an example of use of the algorithm within the class of log-linear models (without a latent variable). We show that, if we restrict ourselves to the subclass of decomposable models, it is possible to implement a simple Gibbs-based MCMC. Some preliminary results on simulated data (not shown in this work for space limitation), seem to indicate that the restriction to that subclass does not affect the efficacy of the procedure.

## 2 The model

Consider  $k$  capturing variables  $\mathbf{Y} = (Y_1, \dots, Y_k)$ , where  $Y_i = 1$  if a certain unit is listed in the  $i$ -th source and 0 otherwise, and let  $X$  be the latent variable taking values in  $\{1, \dots, m\}$  identifying the latent classes of our population. The LCM under the CIA can be equivalently expressed as the mixture model

$$P(\mathbf{Y} = \mathbf{y}) = p_{\mathbf{y}} = \sum_{x=1}^m p_x \prod_{i=1}^k p_{y_i|x}, \quad (1)$$

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

$$[XY_1][XY_2] \cdots [XY_k], \quad (2)$$

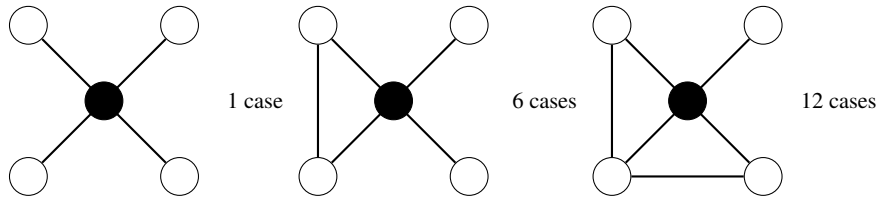
where we use the classic notation reporting only the higher order interactions (generators) of the model. The dependence graph of this model is a star-shaped graph, where the node representing  $X$  is connected to all other nodes, like the one in Figure 1 on the left. Any additional interaction term with respect to (2), (i.e. any additional arc in the graph), constitutes a relaxation of the CIA.

Denote the number of observed units as  $n_{obs}$ , and the number of units presenting the capture profile  $\mathbf{y} \in \{0, 1\}^k$  as  $n_{\mathbf{y}}$ . Let  $n_{x,\mathbf{y}}$  be the number of units having that profile which belong to latent class  $x$ , so that  $\sum_x n_{x,\mathbf{y}} = n_{\mathbf{y}}$ , and  $n_{x,\mathbf{y}}$  be the vector  $(n_{x,\mathbf{y}})_{x=1,\dots,m}$ . Let  $n_0$  be the number of uncaptured units to be estimated and  $N$  the total population count such that  $\sum_{\mathbf{y}} n_{\mathbf{y}} + n_0 = N$ . Let  $M$  be the (random variable associated to) the model to be chosen in a pre determined set  $\mathcal{M}$ , and  $\Theta_M$  the set of parameters associated to model  $M$ .

### 2.1 Prior distributions

Obviously, the choice of a prior on the set of models  $\mathcal{M}$ , as on any other parameter, is subjectively arbitrary. However, we usually just want to exclude some cases, and set a uniform prior on all remaining cases. As we have said, we focus on decomposable models and, as an elementary additional criterion, we rule out all unidentifiable models. In [14] we can find a necessary and sufficient condition for the identifiability of any graphical model (hence of any decomposable model too).

In practice, to utilize the proposed methodology, we have to list out all possible models for a given number of variables. That is, all identifiable models represented by decomposable graphs containing the star-shaped graph relative to the CIA. Consider the case  $k = 4$ : in Figure 1 we have all decomposable graphs grouped by isomorphism. This leaves us with 19 eligible models. When  $k = 5$  the number of identifiable decomposable models goes up to 355.



**Fig. 1** Identifiable decomposable graph models with 4 manifests (empty nodes) and a latent (black node) grouped by classes of isomorphism

As for the prior distributions over the parameters  $\Theta_M$  of each model  $M$ , we utilize the Hyper Dirichlet distribution described in [5]. Such a choice allows us to exploit

the result of [9] giving an analytical formula for the posterior probabilities of the models given the data.

### 3 Model averaging

We analyze the posterior distribution of  $N$  given the observed data  $\{n_{\mathbf{y}}\}$  with  $\mathbf{y} \neq \mathbf{0}$ ,

$$\pi(N | \{n_{\mathbf{y}}\}) = \sum_{M \in \mathcal{M}} \pi(N | M, \{n_{\mathbf{y}}\}) \pi(M | \{n_{\mathbf{y}}\}).$$

The posterior probability of model  $M$  is given by:

$$\pi(M | \{n_{\mathbf{y}}\}) \propto \pi(M) \pi(\{n_{\mathbf{y}}\} | M) = \pi(M) \int \pi(\{n_{\mathbf{y}}\} | M, \Theta_M) \pi(\Theta_M | M) d\Theta_M.$$

In practice, given the computational complexity of calculating the last integral quantity, one can settle for the simplest (first order) approximation of the marginal likelihood of a model based on the Bayesian Information Criteria (BIC), that is,  $\exp(-BIC/2)$ , which, given equal prior probabilities for the models, leads to the following approximation of the weights  $\pi(M | \{n_{\mathbf{y}}\})$ :

$$\frac{\exp(-BIC_M/2)}{\sum_{M \in \mathcal{M}} \exp(-BIC_M/2)}, \quad (3)$$

where  $BIC_M$  is the BIC of model  $M$ . Then, one can use those weights in computing the averaged mean

$$E[N | \{n_{\mathbf{y}}\}] = \sum_{M \in \mathcal{M}} \hat{N}_M \pi(M | \{n_{\mathbf{y}}\}), \quad (4)$$

where  $\hat{N}_M$  is the posterior mean of  $N$  under model  $M$ .

By using formula (4), one should keep in mind that the approximation quality can be poor in some cases, and, in any case, we limit ourselves to a point estimate of  $N$ . A full Bayesian approach to the problem, on the other hand, would result in an estimate of the whole posterior distribution of  $N$  marginalized over  $\mathcal{M}$ .

### 4 The Gibbs sampler

In this section we outline a Gibbs-based MCMC algorithm to sample from the joint distribution of  $(N, N_{\mathbf{x}, \mathbf{y}}, M, \Theta_M)$ , conditioned on the observed data  $\{n_{\mathbf{y}}\}$ . Note that we cannot obtain the full conditionals for all terms: as pointed out in [2] and in [10], given  $n_{\mathbf{x}, \mathbf{0}}$ , the value of  $N$  is deterministically defined. As a workaround, they propose to consider the conditional distribution of the couple  $N, N_{\mathbf{x}, \mathbf{0}}$  conditionally

BMA for capture recapture

on the rest. Similarly, we cannot obtain the conditional distribution of  $M$  given the parameters  $\Theta_M$ . For these reasons, the algorithm loops over the following steps:

- 1) sample  $n_{\mathbf{x},\mathbf{y}}^{(t)}$  from  $\pi(N_{\mathbf{x},\mathbf{y}} | N, M, \Theta_M, \{n_{\mathbf{y}}\})$ , for all  $\mathbf{y} \neq \mathbf{0}$ ,

$$N_{\mathbf{x},\mathbf{y}} \sim \text{Mult}(n_{\mathbf{y}}, p_{\mathbf{x}|\mathbf{y}}),$$

where the  $p_{\mathbf{x}|\mathbf{y}}$  are calculated according to the current value of  $M$  and  $\Theta_M$ ;

- 2) sample a couple  $(N^{(t)}, n_{\mathbf{x},\mathbf{0}}^{(t)})$  from

$$\pi(N, N_{\mathbf{x},\mathbf{0}} | M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\}) = \pi(N | M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\}) \pi(N_{\mathbf{x},\mathbf{0}} | N, M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\}).$$

Note that

$$\pi(N | M, \Theta_M, \{n_{\mathbf{x},\mathbf{y}}\}) = \pi(N | M, \Theta_M, n_{obs}) \propto \pi(N) \binom{N}{n_{obs}} p_{\mathbf{0}}^{N-n_{obs}} (1 - p_{\mathbf{0}})^{n_{obs}},$$

then, if we choose the improper prior  $\pi(N) \propto 1/N$ , the conditional distribution of  $N$  results in a Negative Binomial distribution, and we simply have to

- sample  $N^{(t)}$  from  $\text{NegBin}(n_{obs}, 1 - p_{\mathbf{0}})$ ;
- sample  $n_{\mathbf{x},\mathbf{0}}^{(t)}$  from  $\text{Mult}((N^{(t)} - n_{obs}), p_{\mathbf{x}|\mathbf{0}})$ .

where  $p_{\mathbf{x}|\mathbf{0}}$  and  $p_{\mathbf{0}}$  are calculated according to the current value of  $M$  and  $\Theta_M$ ;

- 3) sample from  $\pi(M, \Theta_M | \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\}) = \pi(\Theta_M | M, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\}) \pi(M | \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\})$ .

That is,

- calculate the posterior probability of each eligible decomposable model and sample  $M^{(t)}$  from  $\pi(M | \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\})$ . The posterior probability of each model  $M$  is defined as a product of Gamma functions (see [9]);
- then sample all parameters  $\Theta_M^{(t)}$  from their posterior conditional distribution  $\pi(\Theta_M | M, \{n_{\mathbf{x},\mathbf{0}}\}, \{n_{\mathbf{x},\mathbf{y}}\})$ , which is a product of Dirichlet distributions.

## 5 Conclusions

The proposed algorithm can be just used for model selection by simply inspecting the generated values of  $M$ , as the relative frequency of each model constitutes an estimate of its posterior probability, and select the best model accordingly. However, model averaging seems to be the best choice in capture-recapture problems. Compared to the usual approximation techniques, our estimates should be more accurate, and allow to inspect the whole posterior distribution of  $N$  at the cost of some additional computational effort which appears nonetheless reasonable. The restriction to decomposable models may seem a severe limiting factor, as they constitutes a minority fraction of the possible models, and many frequently used models, such as the one with all second order and no higher order interactions, are left out. How-



ever, some preliminary results based on simulations (not shown in this work for space limitation), appears to be encouraging. In fact, the proposed approach seems to work well even when used on data generated from non decomposable models.

## References

- [1] A. Agresti. Simple capture–recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, 50(2):494–500, 1994.
- [2] S. Basu and N. Ebrahimi. Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, 88(1):269–279, 2001.
- [3] A. Biggeri, E. Stanghellini, F. Merletti, and M. Marchi. Latent class models for varying catchability and correlation among sources in Capture-Recapture estimation of the size of a human population. *Statistica Applicata*, 11(3):1–14, 1999.
- [4] B.A. Coull and A. Agresti. The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, 55(1):294–301, 1999.
- [5] A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 21(3):1272–1317, 1993.
- [6] D. Di Cecco. Estimating population size in multiple record systems with uncertainty of state identification. In *Analysis of Integrated Data*, pages 169–196. Chapman and Hall/CRC, 2019.
- [7] D. Di Cecco, M. Di Zio, and B. Liseo. Bayesian latent class models for capture–recapture in the presence of missing data. *Biometrical Journal*, 2020.
- [8] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
- [9] D. Madigan and J. C. York. Bayesian methods for estimation of the size of a closed population. *Biometrika*, 84(1):19–31, 1997.
- [10] D. Manrique-Vallier. Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, 72(4):1246–1254, 2016.
- [11] A. M. Overstall and R. King. `conting`: An R package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software*, 58(7):1–27, 2014.
- [12] E. Stanghellini and M. G. Ranalli. Population size estimation using a categorical latent variable. In *Capture-Recapture Methods for the Social and Medical Sciences*, pages 277–290. Chapman and Hall/CRC, 2017.
- [13] E. Stanghellini and P. G. M. van der Heijden. A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account. *Biometrics*, 60(2):510–516, 2004.
- [14] E. Stanghellini and B. Vantaggi. Identification of discrete concentration graph models with one hidden binary variable. *Bernoulli*, 19(5A):1920–1937, 2013.