UNIVERSITÀ DI PISA

Sant'Anna
Scuola Universitaria Superiore Pisa

Consiglio Nazionale delle Ricerche

Book of Short Papers
SIS 2020

SIS 2020
Pisa

SIS
Società
Italiana di
Statistica

Editors: Alessio Pollice, Nicola Salvati and Francesco Schirripa Spagnolo

Copyright © 2020

PUBLISHED BY PEARSON

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
Tiziana Tuoto and Marco Fortini

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
Ilaria Benedetti and Federico Crescenzi

Communication of Uncertainty of Official Statistics. 16
Edwin de Jonge and Gian Luigi Mazzi

Measuring uncertainty for infra-annual macroeconomic statistics. 22
George Kapetanios, Massimiliano Marcellino and Gian Luigi Mazzi

Bayesian methods in biostatistics 27

Network Estimation of Compositional Data. 28
Nathan Osborne, Christine B. Peterson and Marina Vannucci

Using co-data to empower genomics-based prediction and variable selection. 34
Magnus M. Münch, Mirrelijn M. van Nee and Mark A. van de Wiel

Data integration versus privacy protection: a methodological challenge? 40

Statistical Disclosure Control for Integrated Data. 41
Natalie Shlomo

The Integrated System of Statistic Registers: first steps towards facing privacy issues. 47
Mauro Bruno and Roberta Radini

Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics. 53
Fabio Ricciato, Kostas Giannakouris, Albrecht Wirthmann and Martina Hahn

Designing adaptive clinical trials 59

Optimal designs for multi-arm exponential trials. 60
Rosamarie Frieri and Marco Novelli

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
Isabella Sulis, Silvia Columbu and Mariano Porcu

Pull factors for university students' mobility: a gravity model approach. 73
Giovanni Boscaino and Vincenzo Giuseppe Genova

Spatial autoregressive gravity models to explain the university student mobility in Italy. 79
Silvia Bacci, Bruno Bertaccini and Chiara Bocci

Environmental Statistics (GRASPA-SIS) 85

A Time Clustering Model for Spatio-Temporal Data. 86
Clara Grazian, Gianluca Mastrantonio and Enrico Bibbona

Reconstruction of sparsely sampled functional time series using frequency domain functional principal components. 93
Amira Elayouty, Marian Scott and Claire Miller

Methods for High Dimensional Compositional Data Analysis 98

Algorithms for compositional tensors of third-order. 99
Violetta Simonacci

High-dimensional regression with compositional covariates: a robust perspective. 105
Gianna Serafina Monti and Peter Filzmoser

Three-way compositional analysis of energy intensity in manufacturing. 111
Valentin Todorov and Violetta Simonacci

Modern Statistics for Physics Discoveries 117

Identification of high-energy λ-ray sources via nonparametric clustering. 118
Giovanna Menardi, Denise Costantin, and Federico Ferraccioli

Statistical Analysis of Macroseismic Data for a better Evaluation of Earthquakes Attenuation Laws. 124
Marcello Chiodi, Antonino D'Alessandro, Giada Adelfio and Nicoletta D'Angelo

Network Modelling in Biostatistics..130

Natural direct and indirect relative risk for mediation analysis. 131
Monia Lupparelli and Alessandra Mattei

New issues on multivariate and univariate quantile regression.................137

Mixtures of quantile regressions for longitudinal data: an R package. 138
Maria Francesca Marino, Maria Giovanna Ranalli and Marco Alfò

Multivariate Mixed Hidden Markov Model for joint estimation of multiple quantiles. 144
Luca Merlo, Lea Petrella and Nikos Tzavidis

Recent methodological advances in finite mixture modeling
with applications (CLADAG-SIS) 150

Aggregating Gaussian mixture components. 151
Roberto Rocci

Local and overall coefficients of determination for mixtures of generalized linear models. 157
Roberto Di Mari, Salvatore Ingrassia and Antonio Punzo

Statistical Analysis of Satellite Data (SDS-SIS) 163

Functional Data Analysis for Interferometric Syntethic Aperture Radar Data Post-Processing:
The case of Santa Barbara mud volcano. 164
Matteo Fontana, Alessandra Menafoglio, Francesca Cigna and Deodato Tapete

Recent Contributions to the Understanding of the Uncertainty in Upper-Air Reference Measurements. 170
Alessandro Fassò

Statistical models and methods for Business and Industry 176

Modelling and monitoring of complex 3D shapes: a novel approach for lattice structures. 177
Bianca Maria Colosimo, Marco Grasso and Federica Garghetti

Open data powered territorial planning - Case study: The Turin historical center. 183
Silvia Casagrande, Gianmaria Origgi, Alberto Pasanisi, Martina Tamburini, Pascal Terrien, Tania Cerquitelli and Alfonso Capozzoli

Process optimization in Industry 4.0: Are all data analytics models useful? 189
Alberto Ferrer

Technology and demographic behaviours (AISP-SIS) 195

Internet and the Timing of Births. 196
Maria Sironi, Osea Giuntella and Francesco C. Billari

The Internetization of Marriage: Effects of the Diffusion of High-Speed Internet on Marriage, Divorce, and Assortative Mating. 202
Francesco C. Billari, Osea Giuntella and Luca Stella

Solicited Sessions

Advanced Statistical Methods in Health Analytics 209

Assessing the impact of the intermediate event in a non-markovian illness-death model. 210
Davide Paolo Bernasconi, Elena Tassistro, Maria Grazia Valsecchi and Laura Antolini

Big data and AI: challenges and opportunities in healthcare. 216
Vieri Emiliani, Gian Luca Cattani and Fabrizio Selmi

Statistical methodology for volume-outcome studies. 222
Marta Fiocco and Floor van Oudenhoven

Advances in textual data mining 228

Distance measures for exploring pairs of novels in a large corpus of Italian literature. 229
Matilde Trevisani and Arjuna Tuzzi

Supervised vs Unsupervised Latent Dirichlet Allocation: topic detection in lyrics. 235
Mariangela Sciandra, Alessandro Albano and Irene Carola Spera

Advances in the interaction between artificial intelligence and official statistics 241

Automated Land Cover Maps from Satellite Imagery by Deep Learning. 242
Fabrizio De Fausti, Francesco Pugliese and Diego Zardetto

CROWD4SDG: Crowdsourcing for sustainable developments goals. 248
Barbara Pernici

Permanent Population Census: evaluation of the effects of regional strategies on the process efficiency.
The direct experience of Tuscany. 253
Linda Porciani, Luisa Francovich, Luca Faustini and Alessandro Valentini

Capture-recapture methods 259

Bayesian Model Averaging for Latent Class Models in Capture-Recapture. 260
Davide Di Cecco

Combining “signs of life” and survey data through latent class models to consider over-coverage in Capture-Recapture estimates of population counts. 266
Marco Fortini, Antonella Bernardini, Marco Caputi and Nicoletta Cibella

Population size estimation with interval censored counts and external information. 272
Alessio Farcomeni

Changes in environment extremes and their impacts 278

FPCA Clustering of rainfall events. 279
Gianluca Sottile, Antonio Francipane, Leonardo Noto and Giada Adelfio

Trends in rainfall extremes in the Venice lagoon catchment. 285
Ilaria Prosdocimi and Carlo Gaetan

Copulas: models and inference 291

Analysis of district heating demand through different copula-based approaches. 292
F. Marta L. Di Lascio and Andrea Menapace

CoVaR and backtesting: a comparison between a copula approach and parametric models. 298
Michele Leonardo Bianchi, Giovanni De Luca and Giorgia Rivieccio

Estimating Asymmetric Dependence via Empirical Checkerboard Copulas. 304
Wolfgang Trutschnig and Florian Griessenberger

Strong Convergence of Multivariate Maxima. 310
Michael Falk, Simone A. Padoan and Stefano Rizzelli

Data Science: when different expertise meet 316

Bayesian stochastic modelling of the temporal evolution of seismicity. 317
Elisa Varini and Renata Rotondi

Cluster Analysis for the Characterization of Residential Personal Exposure to ELF Magnetic Field. 323
Gabriella Tognola, Silvia Gallucci, Marta Bonato, Emma Chiaramello, Isabelle Magne, Martine Souques, Serena Fiocchi, Marta Parazzini and Paolo Ravazzani

Statistical Assessment and Validation of Ship Response in High Sea State by Computational Fluid Dynamics. 328
Andrea Serani, Matteo Diez and Frederick Stern

Uncertainty Quantification for PDEs with random data using the Multi-Index Stochastic Collocation method. 334
Lorenzo Tamellini and Joakim Beck

Emerging challenges in official statistics:
new data sources and methods 340

Small area poverty indicators adjusted using local spatial price indices. 341
Stefano Marchetti, Luigi Biggeri, Caterina Giusti and Monica Pratesi

Smart solutions for trusted smart statistics: the European big data hackathon experience. 347
Francesco Amato, Mauro Bruno, Tania Cappadozzi, Fabrizio De Fausti and Manuela Michelini

The ESSnet Project Smart Surveys: new data sources and tools for Surveys of Official Statistics 353

Factorial and dimensional reduction methods for the construction
of indicators for evaluation (SVQS-SIS)..359

A comparison of MBC with CLV and PCovR methods for dimensional reduction of the soccer players' performance attributes. 360
Maurizio Carpita, Enrico Ciavolino and Paola Pasca

A framework of cumulated chi-squared type statistics for ordered correspondence analysis. New tools and properties. 366
Antonello D'Ambra, Pietro Amenta and Luigi D'Ambra

Exploring drug consumption via an ultrametric correlation matrix. 372
Giorgia Zaccaria and Maurizio Vichi

Ranking extraction in ordinal multi-indicator systems. 378
Marco Fattore and Alberto Arcagni

Gender statistics 384

Gender differences in Italian STEM degree courses: a discrete-time competing-risks model. 385
Marco Enea and Massimo Attanasio

Some Challenges and Results in Measuring Gender Inequality. 391
Fabio Crescenzi and Francesco Di Pede

How Deep is Your Plot? Young SIS and deep statistical learning (ySIS)..397

A modal approach for clustering matrices. 398
Federico Ferraccioli and Giovanna Menardi

A Note on Detection of Perturbations in Biological Networks. 404
Vera Djordjilović

Bayesian inference for DAG-probit models. 410
Federico Castelletti

Variational Bayes for Gaussian Factor Models under the Cumulative Shrinkage Process. 416
Sirio Legramanti

Measuring poverty and vulnerability 421

Choosing the vulnerability threshold using the ROC curve. 422
Chiara Gigliarano and Conchita D'Ambrosio

New advances in applications, a Bayesian nonparametric perspective 428

Bayesian Mixture Models for Latent Class Analysis. 429
Raffaele Argiento, Bruno Bodin and Maria De Iorio

Non-Parametric Inference and Forecasting of Functional and Object Data 435

An interpretable estimator for the function-on-function linear regression model with application to the Canadian weather data. 436
Fabio Centofanti and Matteo Fontana

Statistical process monitoring of multivariate profiles from ship operating conditions. 440
Christian Capezza

Prior choice in Bayesian Modelling (SISbayes)..446

Bayesian Learning of Multiple Essential Graphs. 447
Luca La Rocca, Federico Castelletti, Stefano Peluso, Francesco Claudio Stingo and Guido Consonni

Bayesian post-processing of Gibbs sampling output for variable selection. 453
Stefano Cabras

Priors on precision parameters of IGRMF models. 459
Aldo Gardini, Fedele Greco and Carlo Trivisano

Sequence Analysis: methods and applications 465

Internal migration, family formation and social stratification in Europe. A life course approach. 466
Roberto Impicciatore, Gabriele Ballarino and Nazareno Panichella

Socio economic integration of migrants 472

A study on the characteristics of spouses who intermarry in Italy. 473
Agnese Vitali and Romina Fraboni

Statistical Analysis for mobility and transportation 479

A multilevel Analysis of University attractiveness in the network flows from Bachelor to Master's degree. 480
Silvia Columbu and Ilaria Primerano

Analysis of mobility data through a novel Cheng and Church algorithm for functional data. 486
Marta Galvani, Agostino Torti and Alessandra Menafoglio

Bridge closures in a transportation network: analysis of the impacts in the region of Lombardy. 491
Agostino Torti, Marika Arena, Giovanni Azzone, and Piercesare Secchi

Statistical Methods and Applications in Social Network Analysis 496

A clustering procedure for ego-networks data: an application to Italian elders living in couple. 497
Elvira Pelle and Roberta Pappadà

Analysing the mediating role of a network: a Bayesian latent space approach. 503
Chiara Di Maria, Antonino Abbruzzo and Gianfranco Lovison

Network-time autoregressive models for valued network panel. 509
Viviana Amati

University student mobility flows and network data structures. 515
Maria Prosperina Vitale, Giuseppe Giordano and Giancarlo Ragozini

Statistical Methods in Psychometrics 521

A simple probabilistic model to evaluate questionable interim analysis strategies. 522
Francesca Freuli and Luigi Lombardi

Incorporating Expert Knowledge in Structural Equation Models: Applications in Psychological Research. 528
Gianmarco Altoè, Claudio Zandonella Callegher, Enrico Toffalini and Massimiliano Pastore

Predicting social media addiction from Instagram profiles: A data mining approach. 534
Antonio Calcagnì, Veronica Cortellazzo, Francesca Guizzo, Paolo Girardi, Natale Canale

Structural entropy based modeling for psychological measurement. 540
Enrico Ciavolino, Mario Angelelli, Paola Pasca and Omar Carlo Gioacchino Gelo

Statistical modelling in environmental epidemiology.................................546

A Time Varying Coefficient Model to Estimate the Short-Term Effects of Air Pollution on Human Health. 547
Pasquale Valentini, Luigi Ippoliti and Clara Grazian

Joint Analysis of Short and Long-Term Effects of Air Pollution. 551
Annibale Biggeri, Dolores Catelan, Giorgia Stoppa and Corrado Lagazio

Statistical Modelling of Scientific Evidence for Forensic Investigation and Interpretation..557

DNA mixtures with related contributors. 558
Peter J. Green and Julia Mortera

Forensic Statistics: How to estimate life expectancy after injury. 564
Jane L Hutton

The additional contribution of combining genetic evidence from multiple samples in a complex case. 570
Giampietro Lago

The history of forensic inference and statistics: a thematic perspective. 576
Franco Taroni and Colin Aitken

Topological learning: interpretable representations of complex data......581

Comparing Neural Networks via Generalized Persistence. 582
Mattia G. Bergomi and Pietro Vertechi

On the topological complexity of decision boundaries. 588
António Leitão and Giovanni Petri

Persistence-based Kernels for Data Classification. 594
Ulderico Fugacci

Topological and Mixed-type learning of Brain Activity. 600
Tullia Padellini, Pierpaolo Brutti, Riccardo Giubilei

Contributed papers and Posters

Bayesian Statistics 607

A Bayesian approach for modelling dependence among mixture densities. 608
Mario Beraha, Matteo Pegoraro, Riccardo Peli and Alessandra Guglielmi

A change of glasses strategy to solve the rare type match problem. 614
Giulia Cereda and Fabio Corradi

A new prior distribution on the simplex: the extended flexible Dirichlet. 620
Roberto Ascari, Sonia Migliorati and Andrea Ongaro

ABC model choice via mixture weight estimation. 626
Gianmarco Caruso, Luca Tardella and Christian P. Robert

An ABC algorithm for random partitions arising from the Dirichlet process. 632
Mario Beraha and Riccardo Corradin

Bayesian Inference of Undirected Graphical Models from Count Data. 638
Pier Giovanni Bissiri, Monica Chiogna and Nguyen Thi Kim Hue

Bayesian IRT models in NIMBLE. 644
Sally Paganin, Chris Paciorek and Perry de Valpine

Bayesian modelling of Facebook communities via latent factor models. 650
Emanuele Aliverti

Bayesian nonparametric adaptive classification with robust prior information. 655
Francesco Denti, Andrea Cappozzo and Francesca Greselin

Choosing the right tool for the job: a systematic analysis of general purpose MCMC software. 661
Mario Beraha, Giulia Gualtieri, Eugenia Villa, Riccardo Vitali and Alessandra Guglielmi

Empirical Bayes estimation for mixture models. 667
Catia Scricciolo

Improving ABC via Large Deviations Theory. 673
Cecilia Viscardi, Michele Boreale and Fabio Corradi

Learning Bayesian Networks for Nonparanormal Data. 679
Flaminia Musella and Vincenzina Vitale

Measuring well-being combining different data sources: a Bayesian networks approach. 685
Federica Cugnata, Silvia Salini and Elena Siletti

Penalising the complexity of extensions of the Gaussian distribution. 691
Diego Battagliese and Brunero Liseo

Predictive discrepancy of credible intervals for the parameter of the Rayleigh distribution. 697
Fulvio De Santis and Stefania Gubbiotti

Small-area statistical estimation of claim risk. 702
Francesca Fortunato, Fedele Greco and Pierpaolo Cristaudo

Subject-specific Bayesian Hierarchical model for compositional data analysis. 708
Matteo Pedone and Francesco C. Stingo

Wasserstein consensus for Bayesian sample size determination. 714
Michele Cianfriglia, Tullia Padellini and Pierpaolo Brutti

Biostatistics 720

A comparison of the CAR and DAGAR spatial random effects models with an application to diabetics rate estimation in Belgium. 721
Vittoria La Serra, Christel Faes, Niel Hens and Pierpaolo Brutti

A functional approach to study the relationship between dynamic covariates and survival outcomes: an application to a randomized clinical trial on osteosarcoma. 727
Marta Spreafico, Francesca Ieva and Marta Fiocco

A Statistical Approach to the Alignment of fMRI Data. 733
Angela Andreella, Ma Feilong, Yaroslav Halchenko, James Haxby and Livio Finos

Adaptive clinical trials: Bayesian decision-theoretic and frequentist approaches for cost-effectiveness analysis. 739
Martin Forster and Marco Novelli

Bootstrap corrected Propensity Score: Application for Anticoagulant Therapy in Haemodialysis Patients. 745
Maeregu W. Arisido, Fulvia Mecatti and Paola Rebora

Combining multiple sources to overcome misclassification bias in epidemiological database studies. 751
Francesca Beraldi, Rosa Gini, Emanuela Dreassi, Leonardo Grilli and Carla Rampichini

Deep Sparse Autoencoder-based Feature Selection for SNPs Validation in Prostate Cancer Radiogenomics. 756
Michela Carlotta Massi, Francesca Ieva, Anna Maria Paganoni, Andrea Manzoni, Paolo Zunino, Nicola Rares Franco, Tiziana Rancati and Catharine West

Graphical models for count data: an application to single-cell RNA sequencing. 762
Nguyen Thi Kim Hue, Monica Chiogna and Davide Risso

Interregional mobility, socio-economic inequality and mortality among cancer patients. 768
Claudio Rubino, Mauro Ferrante, Antonino Abbruzzo, Giovanna Fantaci and Salvatore Scondotto

PET radiomics-based lesions representation in Hodgkin lymphoma patients. 774
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

Prediction of late radiotherapy toxicity in prostate cancer patients via joint analysis of SNPs sequences. 780
Nicola Rares Franco, Michela Carlotta Massi, Francesca Ieva, Anna Maria Paganoni, Andrea Manzoni, Paolo Zunino, Tiziana Rancati and Catharine West

Predictive versus posterior probabilities for phase II trial monitoring. 785
Valeria Sambucini

Profile networks for precision medicine. 791
Andrea Lazzerini, Monia Lupparelli and Francesco C. Stingo

Proton-Pump Inhibitor Provider Profiling via Funnel Plots and Poisson Regression. 797
Dario Delle Vedove, Francesca Ieva and Anna Maria Paganoni

Selecting optimal thresholds in ROC analysis with clustered data. 803
Duc Khanh To, Gianfranco Adimari and Monica Chiogna

Environment, Physics and Engineering 809

A hidden semi-Markov model for segmenting environmental toroidal data. 810
Francesco Lagona and Antonello Maruotti

An experimental analysis on quality and security about green communication. 816
Vito Santarcangelo, Emilio Massa, Davide Scintu, Michele Di Lecce and Massimiliano Giacalone

An improved sensitivity-data based method for probabilistic ecological risk assessment. 822
Sonia Migliorati and Gianna Serafina Monti

Comparing predictive distributions in EMOS. 828
Giummolè Federica and Mameli Valentina

Compositional analysis of fish communities in a fast changing marine ecosystem. 834
Pierfrancesco Alaimo Di Loro, Marco Mingione, Giovanna Jona Lasinio, Sara Martino and Francesco Colloca

FDA dimension reduction techniques and components separation in Fourier-transform infrared spectroscopy. 840
Francesca Di Salvo, Elena Piacenza and Delia Francesca Chillura Martino

Functional Data Analysis for Spectroscopy Data. 846
Mara S. Bernardi, Matteo Fontana, Alessandra Menafoglio, Diego Perugini, Alessandro Pisello, Marco Ferrari, Simone De Angelis, Maria Cristina De Sanctis and Simone Vantini

Functional graphical model for spectrometric data analysis. 852
Laura Codazzi, Alessandro Colombi, Matteo Gianella, Raffaele Argiento, Lucia Paci and Alessia Pini

Local LGCP estimation for spatial seismic processes. 857
Nicoletta D'Angelo, Marianna Siino, Antonino D'Alessandro and Giada Adelfio

Observation-driven models for storm counts. 863
Mirko Armillotta, Alessandra Luati and Monia Lupparelli

Statistical control of complex geometries, with application to Additive Manufacturing. 869
Riccardo Scimone, Tommaso Taormina, Bianca Maria Colosimo, Marco Grasso, Alessandra Menafoglio, Piercesare Secchi

Tree attributes map by 3P sampling in a design-based framework. 875
Lorenzo Fattorini and Sara Franceschi

Unsupervised classification of texture images by gray-level spatial dependence matrices and genetic algorithms. 880
Roberto Baragona and Laura Bocci

Finance, business and official statistics 886

A discrete choice approach to analyze contractual attributes in the durum wheat sector in Italy. 887
Stefano Ciliberti, Simone Del Sarto, Giulia Pastorelli, Angelo Frascarelli and Gaetano Martino

A fuzzy approach to the measurement of the employment rate. 893
Bruno Cheli, Alessandra Coli and Andrea Regoli

A proposal to model credit risk contagion using network count-based models. 898
Arianna Agosto and Daniel Felix Ahelegbey

A similarity matrix approach to empower ESCO interfaces for testing, debugging and in support of users' experience. 904
Adham Kahlawi, Cristina Martelli, Lucia Buzzigoli, Laura Grassini

Adding MIDAS terms to Linear ARCH models in a Quantile Regression framework. 910
Vincenzo Candila and Lea Petrella

Company requirements in Italian tourism sector: an analysis for profiles. 916
Paolo Mariani, Andrea Marletta, Lucio Masserini and Mariangela Zenga

Determinants of Firms' Default Risk after the 2008 and 2011 Economic Crises: a Latent Growth Models Approach. 921
Lucio Masserini, Matilde Bini and Alessandro Zeli

Double Asymmetric GARCH-MIDAS model - new insights and results. 927
Alessandra Amendola, Vincenzo Candila and Giampiero M. Gallo

European SMEs and Circular Economy Activities: Evaluating the Advantage on Firm Performance through the Estimation of Average Treatment Effects. 933
Luca Secondi

Financial Spillover Measures to Assess the Stability of Basket-based Stablecoins. 939
Paolo Pagnottoni

Forecasting Banknote Flows in BdI Branches: Speed-up with Machine Learning. 945
Marco Brandi, Monica Fusaro, Tiziana Laureti and Giorgia Rocco

Fully reconciled GDP forecasts from Income and Expenditure sides. 951
Luisa Bisaglia, Tommaso Di Fonzo and Daniele Girolimetto

GLASSO Estimation of Commodity Risks. 957
Beatrice Foroni, Saverio Mazza, Giacomo Morelli and Lea Petrella

Measuring the Effect of Unconventional Policies on Stock Market Volatility. 963
Giampiero M. Gallo, Demetrio Lacava and Edoardo Otranto

Multidimensional versus unidimensional poverty measurement. 969
Michele Costa

Multiple outcome analysis of European Agriculture in 2000-2016: a latent class multivariate trajectory approach. 975
Alessandro Magrini

Nowcasting GDP using mixed-frequency based composite confidence indicators. 981
Maria Carannante, Raffaele Mattera, Michelangelo Misuraca, Germana Scepi and Maria Spano

On the tangible and intangible assets of Initial Coin Offerings. 987
Paola Cerchiello and Anca Mirela Toma

Seasonality variation of electricity demand: decompositions and tests. 993
Luigi Grossi and Mauro Mussini

SMEs circular economy practices in the European Union: Implications for sustainability. 999
Nunzio Tritto, Josè G. Dias and Francesca Bassi

Tax Incentives' Effect on the Provision of Occupational Welfare in Italian Enterprises. 1005
Alessandra Righi

The determinants of eco-innovation: a country comparison using the community innovation survey. 1011
Ida D'Attoma and Silvia Pacei

World ranking of urban sustainability through composite indicators. 1017
Elena Grimaccia, Alessia Naccarato and Silvia Terzi

Machine Learning and Data Science..1023

A novel approach for Artificial Intelligence through Lorenz zonoids and Shapley Values. 1024
Paolo Giudici and Emanuela Raffinetti

A warning signal for variable importance interpretation in tree-based algorithms. 1030
Anna Gottard and Giulia Vannucci

Assessment of the effectiveness of digital flyers: analysis of viewing behavior using eye tracking. 1036
Gianpaolo Zammarchi, Claudio Conversano and Francesco Mola

At risk mental status analysis: a comparison of model selection methods for ordinal target variable. 1042
Elena Ballante, Silvia Molteni, Martina Mensi and Silvia Figini

Categorical Encoding for Machine Learning. 1048
Agostino Di Ciaccio

Dynamic Quantile Regression Forest. 1054
Mila Andreani and Lea Petrella

Estimating the UK Sentiment Using Twitter. 1059
Stephan Schlosser, Daniele Toninelli and Michela Cameletti

Forecasting local rice prices from crowdsourced data in Nigeria. 1065
Ilaria Lucrezia Amerise and Gloria Solano Hermosilla

Generalized Mixed Effects Random Forest: does Machine Learning help in predicting university student dropout? 1071
Massimo Pellagatti, Chiara Masci, Francesca Ieva and Anna Maria Paganoni

HateViz: a textual dashboard Twitter data-driven. 1077
Emma Zavarrone, Maria Gabriella Grassia, Marina Marino, Rocco Mazza and Nicola Canestrari

How to perform cyber risk assessment via cumulative logit models. 1083
Silvia Facchinetti, Silvia Angela Osmetti and Claudia Tarantola

Machine learning prediction for accounting system. 1087
Chiara Bardelli and Silvia Figini

Teaching statistics: an assessment framework based on Multidimensional IRT and Knowledge Space Theory. 1093
Cristina Davino, Rosa Fabbricatore, Carla Galluccio, Daniela Pacella, Domenico Vistocco, Francesco Palumbo

The weight of words: textual data versus sentiment analysis in stock returns prediction. 1099
Riccardo Ferretti and Andrea Sciandra

Unsupervised Energy Trees: clustering with complex and mixed-type variables. 1105
Riccardo Giubilei, Tullia Padellini and Pierpaolo Brutti

Using anchoring vignettes to adjust self-reported life satisfaction: a nonparametric approach leading to a Semantic Differential scale. 1111
Sara Garbin, Serena Berretta, Maria Iannario and Omar Paccagnella

Variable selection for robust model-based learning from contaminated data. 1117
Andrea Cappozzo, Francesca Greselin and Thomas Brendan Murphy

Variable Selection in Text Regressions: Back to Lasso? 1123
Marzia Freo and Alessandra Luati

Web Usage Mining and Website Effectiveness. 1129
Maria Francesca Cracolici and Furio Urso

Models and methods - Categorical, Ordinal, Rank Data 1135

Aberration for the analysis of two-way contingency tables. 1136
Roberto Fontana and Fabio Rapallo

An investigation of the paradoxical behaviour of κ-type inter-rater agreement coefficients for nominal data. 1142
Amalia Vanacore and Maria Sole Pellegrino

Analyzing faking-good response data: Combination of a Replacement and a Binomial (CRB) distribution approach. 1148
Luigi Lombardi and Antonio Calcagnì

BOD – min range: A Robustness Analysis Method for Composite Indicators. 1154
Emiliano Seri, Leonardo Salvatore Alaimo and Vittoria Carolina Malpassuti

Comparing classifiers for ordinal variables. 1160
Silvia Golia and Maurizio Carpita

Discovering Interaction Effects Between Subject-Specific Covariates: A New Probabilistic Approach For Preference Data. 1166
Alessio Baldassarre, Claudio Conversano, Antonio D'Ambrosio, Mark De Rooij and Elise Dusseldorp

Hybrid random forests for ordinal data. 1171
Rosaria Simone and Gerhard Tutz

Model-based approach to biclustering ordinal data. 1177
Monia Ranalli and Francesca Martella

New algorithms and goodness-of-fit diagnostics for ranked data modelling with the Extended Plackett-Luce distribution. 1183
Cristina Mollica and Luca Tardella

Non-metric unfolding on augmented data matrix: a copula-based approach. 1189
Marta Nai Ruscone and Antonio D'Ambrosio

Ordinal probability effect measures for dyadic analysis in cumulative models. 1194
Maria Iannario and Domenico Vistocco

Simulated annealing for maximum rater agreement. 1200
Fabio Rapallo and Maria Piera Rogantin

Models and methods – Regression..1206

A Clusterwise regression method for Distributional-valued Data. 1207
Rosanna Verde, Francisco de A. T. de Carvalho and Antonio Balzanella

A nonparametric approach for nonlinear variable screening in high-dimensions. 1213
Francesco Giordano, Sara Milito and Lucia Maria Parrella

Adjusted scores for inference in negative binomial regression. 1219
Euloge C. Kenne Pagui, Alessandra Salvan and Nicola Sartori

Estimation of the treatment effect variance in a difference-in-differences framework. 1224
Marco Doretti and Giorgio E. Montanari

Exploring multicollinearity in quantile regression. 1230
Cristina Davino, Tormod Naes, Rosaria Romano and Domenico Vistocco

Generalized M-quantile random effects model. 1236
Francesco Schirripa Spagnolo and Vincenzo Mauro

Goodness-of-fit assessment in linear quantile regression. 1242
Ilaria Lucrezia Amerise and Agostino Tarsitano

Joint Redundancy Analysis by a multivariate linear predictor. 1248
Laura Marcis and Renato Salvatore

M-quantile regression shrinkage and selection via the lasso. 1254
M. Giovanna Ranalli, Lea Petrella and Francesco Pantalone

New insights into the Conditioning and Gain Score approaches in multilevel analysis. 1260
Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto and Carla Rampichini

Simultaneous confidence regions and curvature measures in nonlinear models. 1265
Claudia Furlan and Cinzia Mortarino

Models and methods – Sampling 1271

Design-based consistency of the Horvitz-Thompson estimator for spatial populations. 1272
Lorenzo Fattorini, Marzia Marcheselli, Caterina Pisani and Luca Pratelli

Empirical likelihood in the statistical matching for informative samples. 1278
Daniela Marella and Danny Pfeffermann

Evaluating a Hybrid One-Staged Snowball Sampling through Bootstrap Method on a Simulated Population. 1284
Venera Tomaselli and Giulio Giacomo Cantone

How optimal subsampling depends on guessed parameter values. 1290
Laura Deldossi and Chiara Tommasi

Indicators for risk of selection bias in non-probability samples. 1296
Emilia Rocco and Alessandra Petrucci

On the behaviour of the maximum likelihood estimator for exponential models under a fixed and a two-stage design. 1302
Caterina May and Chiara Tommasi

Pseudo-population based resamplings for two-stage design. 1308
Pier Luigi Conti, Daniela Marella and Vincenzina Vitale

Models and methods - Theoretical Issues in Statistical Inference 1314

A new mixture model for three-way data. 1315
Salvatore D. Tomarchio, Antonio Punzo and Luca Bagnato

A Sequential Test for the Cpmk Index. 1320
Michele Scagliarini

Probability Interpretations and the Selection of the Most Effective Statistics Method. 1326
Paolo Rocchi

Robust Composite Inference. 1332
Valentina Mameli, Monica Musio, Erlis Ruli and Laura Ventura

Statistical hypothesis testing within the Generalized Error Distribution: Comparing the behavior of some nonparametric techniques. 1338
Massimiliano Giacalone and Demetrio Panarello

Stochastic dependence with discrete copulas. 1344
Fabrizio Durante and Elisa Perrone

Models and methods - Time Series and Longitudinal Data.....................1350

Bootstrap test in Poisson–INAR models. 1351
Lucio Palazzo and Riccardo Ievoli

Continuous Time-Interaction Processes for Population Size Estimation. 1357
Linda Altieri, Alessio Farcomeni, Danilo Alunni Fegatelli and Francesco Palini

Longitudinal data analysis using PLS-PM approach. 1363
Rosanna Cataldo, Corrado Crocetta, Maria Gabriella Grassia and Marina Marino

Long-memory models for count time series. 1369
Luisa Bisaglia, Massimiliano Caporin and Matteo Grigoletto

Combining multiple frequencies in Realized GARCH models. 1375
Antonio Naimoli and Giuseppe Storti

Models with Time-Varying Parameters for Realized Covariance. 1381
Luc Bauwens and Edoardo Otranto

Pitman-Yor mixture models for survival data stratification. 1387
Riccardo Corradin, Luis Enrique Nieto Barajas and Bernardo Nipoti

Prediction is not everything, but everything is prediction. 1393
Leonardo Egidi

The Generalized Dynamic Mixtures of Factor Analyzers for clustering multivariate longitudinal data. 1399
Francesca Martella, Antonello Maruotti and Francesco Tursini

Trends and long-run relations in cointegrated time series observed with noise. 1405
Angelica Gianfreda, Paolo Maranzano, Lucia Parisio and Matteo Pelagatti

Population and society 1411

A dimensionality assessment of refugees' vulnerability through an Item Response Theory approach. 1412
Simone Del Sarto, Michela Gnaldi, Yara Maasri and Edouard Legoupil

Accounting for Interdependent Risks in Vulnerability Assessment of Refugees. 1418
Daria Mendola, Anna Maria Parroco and Paolo Li Donni

Active ageing in China: What are the domains that most affect life satisfaction in the elderly? 1424
Ilaria Rocco

Analyzing the waiting time of academic publications: a survival model. 1430
Francesca De Battisti, Giuseppe Gerardi, Giancarlo Manzi and Francesco Porro

Clustering of food choices in a large sample of students using university canteen. 1436
Valentina Lorenzoni, Isotta Triulzi, Irene Martinucci, Letizia Toncelli, Michela Natilli and Roberto Barale, Giuseppe Turchetti

Cruise passengers' expenditure at destinations: Review of survey techniques and data collection. 1442
Caterina Sciortino, Stefano De Cantis, Mauro Ferrante and Szilvia Gyimóthy

Educational integration of foreign citizen children in Italy: a synthetic indicator. 1448
Alessio Buonomo, Stefania Capecchi and Rosaria Simone

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. 1454
Giorgio Calzolari, Maria Gabriella Campolo, Antonino Di Pino and Laura Magazzini

First and Second Year Careers of STEM Students in Italy: A Geographical Perspective. 1460
Antonella D'Agostino, Giulio Ghellini and Gabriele Lombardi

Future Scenarios and Support Interventions for the Family: Involving Experts' Participation through a Mixed-Method Research Study. 1466
Mario Bolzan, Simone Di Zio, Manuela Scioni and Morena Tartari

Gender and Monetary Policy Preferences: a Diff-in-Diff Approach. 1472
Donata Favaro, Anna Giraldo and Ina Golikja

Headcount based indicators and functions to evaluate the effectiveness of Italian university education. 1478
Silvia Terzi and Francesca Petrarca

Identify the speech code through statistics: a data-driven approach. 1484
Andrea Briglia, Massimo Mucciardi and Jérémi Sauvage

Inspecting cause-specific mortality curves by simplicial functional data analysis. 1490
Marco Stefanucci and Stefano Mazzuco

Intertemporal decision making and childless couples. 1495
Daniela Bellani, Bruno Arpino and Daniele Vignoli

Italian Households' Material Deprivation: Multi-Objective Genetic Algorithm approach for categorical variables. 1501
Laura Bocci and Isabella Mingo

LI-CoD Model. From Lifespan Inequality to Causes of Death. 1507
Andrea Nigri and Susanna Levantesi

Modeling Well-Being through PLS-SEM and K-M. 1513
Venera Tomaselli, Mario Fordellone and Maurizio Vichi

News life-cycle: a multiblock approach to the study of information. 1519
Rosanna Cataldo, Marco Del Mastro, Maria Gabriella Grassia, Marina Marino and Rocco Mazza

Short-term rentals in a tourist town. 1525
Silvia Bacci, Bruno Bertaccini, Gianni Dugheri, Paolo Galli, Antonio Giusti and Veronica Sula

SportIstat: a playful activity to developing statistical literacy. 1531
Alessandro Valentini and Francesca Paradisi

Statistical modeling for some features of Airbnb activity. 1537
Giulia Contu and Luca Frigau

Tertiary students with migrant background: evidence from a cohort enrolled at Sapienza University. 1543
Cristina Giudici, Donatella Vicar and Eleonora Trappolini

The Causal Effect of Immigraton Policies on Income Inequality. 1549
Irene Crimaldi, Laura Forastiere, Fabrizia Mealli and Costanza Tortù

The job condition of academic graduates: a joint longitudinal analysis of AlmaLaurea
and Mandatory Notices of the Ministry of Labour. 1557
Maria Veronica Dorgali, Silvia Bacci, Bruno Bertaccini and Alessandra Petrucci

The joint effect of childcare services and flexible female employment on fertility rate in Europe. 1565
Viviana Cocuccio and Massimo Mucciardi

The Left Behind Generation: How the current Early School Leavers affect tomorrow's NEETs? 1571
Giovanni De Luca, Paolo Mazzocchi, Claudio Quintano and Antonella Rocca

The probability to be employed of young adults of foreign origin. 1577
Alessio Buonomo, Francesca Di Iorio and Salvatore Strozza

The risk of inappropriateness in geriatric wards: a comparison among the Italian regions. 1583
Paolo Mariani, Andrea Marletta, Marcella Mazzoleni and Mariangela Zenga

The role of the accumulation of poverty and unemployment for health disadvantages. 1589
Annalisa Busetta, Daria Mendola, Emanuela Struffolino and Zachary Van Winkle

Unemployment and fertility in Italy. A regional level data panel analysis. 1595
Gabriele Ruiu and Marco Breschi

University drop out and mobility in Italy. First evidences on first level degrees. 1601
Nicola Tedesco and Luisa Salaris

Worthiness-based Scale Quantifying. 1607
Giulio D'Epifanio

Young people in Southern Italy and the phenomenon of immigration: what is their perception? 1613
Nunziata Ribecco, Angela Maria D'Uggento and Angela Labarile

Evaluating a Hybrid One-Staged Snowball Sampling through Bootstrap Method on a Simulated Population

Campionamento a valanga ibrido ad uno stadio: una valutazione con metodo bootstrap su una popolazione simulata

Venera Tomaselli[1], Giulio Giacomo Cantone[2]

Abstract: Snowball sampling is a design where a number of individuals is surveyed in a population and then is requested to share the survey tool among their 'social links'. The aim is to recruit other people into the sample. One-stage recruitment means that individuals who are recruited by people already participant are not requested to recruit any others. Although surveys adopting snowball are financially less expensive than alternatives, the reliability of asymptotical estimates through this method is often questioned. A hybrid snowball sampling is designed such that a quota is randomly sampled, and another through snowball. Through a simulation of a demographic setup we found that bootstrap statistics of a one-staged hybrid design asymptotically show no significant difference to random design. This does not extend into situations where the random quota is smaller. We conclude that more complex setups will be needed before to generalise these results.

Abstract: *Il campionamento a valanga è un disegno campionario dove un sondaggio è sottoposto ad certo numero di individui, ed è poi chiesto a questi di condividere il sondaggio tra i loro 'legami sociali' al fine di reclutarne altri nel campione. Il reclutamento ad un livello prevede che chi è reclutato da chi è già partecipante non ne recluti altri. Sebbene i sondaggi a valanga siano meno costosi delle alternative, l'affidabilità delle stime* di questo metodo è spesso posta in dubbio. Un campionamento a valanga ibrido è un disegno campionario dove una frazione del campione è estratta a sorte ed un'altra reclutata a valanga. Per mezzo di una simulazione di un setup demografico abbiamo trovato che le statistiche bootstrap di un disegno ibrido ad uno stadio non mostrano differenze significative rispetto all'alternativa casuale. Ciò non si estende ad una situazione con una minor frazione di individui estratti a caso. Riteniamo che setup più complessi siano necessari prima di generalizzare questi risultati.*

Key words: hybrid one-staged snowball sampling, bootstrapping, simulated population.

1 Introduction to snowball sampling

The terminology 'snowball sampling' denotes the methodology of social research where "a small [random] sample of persons [is interviewed], asking who their best friends are, interviewing these friends, then asking them their friends, interviewing these, and so on" (Coleman, 1958, 29). Leo Goodman (1961), who was a colleague of Coleman at University of Chicago, adopted "snowball sampling" to refer to a

[1] Venera Tomaselli, Department of Political and Social Sciences, University of Catania IT, e-mail: venera.tomaselli@unict.it (*corresponding author*).

[2] Giulio Giacomo Cantone, Department of Physics and Astronomy, University of Catania IT, e-mail: prgcan@gmail.com.

mathematical model to “make statistical inferences about various aspects of the relationships present in the population” by “data obtained using an s stage k name snowball sampling procedure” (*idem*, 148).

The aim of Goodman was to formalize conditions such to assume asymptotically unbiased estimates from Coleman’s procedure. Goodman’s model assumes that k, the amount of interviewed people per s stages of the chain, is a constant. “Stage 0” or $s = 0$ is the initial sample of k cases in the population and is random, instead. Snijders (1992) offered a different formal model with no fixed k. Both authors conclude that while a snowball procedure is not a fully randomized method, if ‘homophily’ among participants is sufficiently low, the sample may be assumed to be representative of a population.

Homophily is a terminology introduced since the ’50 in sociological works and later formalized by network scientists as a parametrization of the correlation between the value of the variable in a node and the amount of its edges (Newman, 2010). Nevertheless, the interpretation of homophily is often ambiguous. Crawford, Aronow, Zeng et al. (2017) claim that homophily is often confused with “preferential recruitment” which is when the probability to be drawn in the sample from stage 1 cannot be assumed uniformly distributed.

When Goodman (2011) returned on the topic of snowball sampling, he highlighted an issue: while the model he developed assumed a random primary sample (‘stage 0’), many studies performed after lacked this assumption[3]. However, Craig, Hays, Pickard et al. (2013) found that on 7 surveyed panel vendors, 6 had a mean of 20% of their proposed participants involved in at least one other of the 6. This leads to intuitively think that samples can support a certain ‘quota’ of biased cases within, without a great loss of precision in estimates within a multivariate design.

The research of Etter and Perneger (2000) is very noteworthy in this sense. Authors surveyed a random sample of 1000 residents in Geneva aged 18-70 (primary participants) in 1997. They asked every contacted subject, even those not willing to participate, to mail the questionnaire to all the smoker and ex-smoker residents in Geneva they knew (secondary participants). At the end of the data gathering process, 3,300 residents were mailed with the questionnaire and 1,167 individuals (35%) returned the questionnaire filled. Of these, 578 were primary participants and 566 were secondary participants. According to authors the mean age difference between the samples was only of 1.7 years (p-value=0.003). The only other significant difference was in sex ratio (7% difference, p-value=0.009) while behavioural traits were reported to be not significantly different in the two groups. As unintended consequence, the authors obtained two samples very similar in size within the one-staged snowball. We call this sampling design ‘hybrid one-staged snowball sampling’.

In a study comparing a random statistical approach to a snowball-based approach to estimate morbidity and mortality of the rare disease visceral Leishmaniosis in two districts of the state of Bihar, India in 2011, Siddiqui, Rabidas, Sinha et al. (2016) concluded that snowball approach was found not sensitive enough to be adopted to estimate morbidity of the rare disease. At the same time, authors noticed that comparing costs, snowball approach required 1/6 of man-days and half the financial costs of the random alternative. Since full snowball design could not be suited for many applications, we want to evaluate if adoption of hybrid sampling can overcome losses of accuracy while reducing costs of research.

2 Simulation of a hybrid sampling

Testing procedures on empirical data in population studies raises the following issues:

- usually empirical research is designed for finding information over a phenomenon, not to provide data for evaluative research about methods
- we lack ‘true values’ on parameters of variables. We can test the hypothesis if two samples are drawn from the same population within confidence level (CL) but this says nothing about which design is better performing on estimation.

We propose instead a combination of two computational methods:

[3] According to the author, this was a result of historical label of ‘snowball’ mostly for the Coleman’s procedure of data collection. Hence, the technique (also referred as ‘chain-referral sampling’) was associated to qualitative studies on ‘rare population’ or on ‘hidden-populations’ (Atkinson and Flint, 2001). Erikson (1979) and Snijders (1992) expressed concern about the actual possibility to randomly sample the stage 0 for hidden populations.

- a simulation model to procedurally generate a virtual dataset of population data from known parameters. Simulated populations can be modeled to fit pre-existing data with a small error (Alfons, Kraft, Templ et al., 2011; Burgard, Kolb, Murkle et al., 2017) or according with a theoretical model i.e. to construct hypotheses
- computational *bootstrap* sampling, falling under the more generic terminology of Monte Carlo methodologies (Gil, Montenegro, González-Rodríguez et al., 2006; Gobet, 2016).

We want to test performance of a sampling design against a standard. Hence, the simulated population does not need to fit empirical data. Still, this virtual population needs to be generated according to non-unrealistic assumptions over expected general outcomes. In particular, assumptions will take form of structural equations made through random variables. Technical sophistications in assumptions of the model may be added once the validity of the simplest models is established.

Bootstrap means that the model is then run many times, and each time estimates are recorded. Hypotheses are tested on the statistics (e.g. the so-called bootstrap-mean) of the distribution of values of the recorded estimates.

The general issues we found for the proposed simulation are two:

Issue 1: How can we simulate non-unrealistic variables in a virtual population?

In order to reflect the complexity of population studies without losing simplicity of linear structural equations, N simulated individual in the population (or 'agent' of the model) will be designed such to show few random variables:

- age: X_0; Gompertz and Weibull are established density functions to model age (Ricklefs and Scheuerlein, 2002). As the Weibull's function has only two parameters, *scale* (λ) and *slope* (k), we adopt it because it is a simpler assumption for our purposes[4].We propose:

$$X_0 \sim Weibull\left(\lambda_{X_0} = 50, k_{X_0} = 3\right). \qquad [1]$$

  For ethical issues, data on minors are usually not collected; therefore, any agent with $X_0 < 18$ (~ 4,56% of N) will be removed from the virtual population.
- sex: X_1; a dummy variable with a probability equal to 0.5 per side:

$$p(X_1 = 1) = 0.5 \qquad [2]$$

- behaviour: X_2; the research on behavioural metrics is a vast field with many insightful contributions on proper tools to measure behaviour's presence and extension (Kline, 2015). But we aim at keeping unambiguity and simplicity in assumptions before prompting actual 'field knowledge' into it. Therefore, we will just notice the presence/absence of the behaviour into a dummy variable with a probability density. For the proposal of the simulation, we arbitrarily set again 0.5 per side:

$$p(X_2 = 1) = 0.5 \qquad [3]$$

- chronic condition: Y; the outcome of a structural equation where X_0 of the agent is the baseline, X_1 and X_2 are endogenous factors, β are weights, and ε is a continuous random error variable which sums all exogenous factors of Y:

$$Y = \begin{cases} 0, X_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon < t_\zeta \\ 1, X_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon > t_\zeta \end{cases} \qquad [4]$$

t is an integer threshold value such that the probability of Y=0 in the whole population asymptotically tends to ζ. To intuitively understand the equation: if we keep out the agents when $X_0 < 18$, set β to be equals to 0 and ε is symmetrical with an expected value equal to 0, then, for ζ =0.5, the value 45 approximates $t_{\zeta=0.5}$. In simple terms, 45 will be the median age of the simulated population, given these parameters.

We can use this information to model our assumptions on β. If we assume:

- normal distribution of ε with μ=0
- age, sex, and behaviour having all equal impact on chronic condition
- $\beta_1 = \beta_2 = 90$

then $t_{\zeta=0.5} = 135$[5].

The threshold t can be set equal to 0 by the property of linear transformation of normal distributions (Bryc, 1995). Once -(t_ζ) is added to μ_ε, we notice that the lower the μ_ε, the higher the ζ. Therefore, we

[4] The Gompertz's assumptions about the parameters may be more suited for fitting synthetic data, instead.

[5] $\boldsymbol{E}(x_0) + \boldsymbol{E}(\beta_1 x_1) + \boldsymbol{E}(\beta_1 x_2) = t_{\zeta=0.5}$.

computationally find a setup pair of values for μ_ε and σ_ε such that the expected value of ζ is approximates a desired value. We want a model where the chronic condition afflicts 1 agent every 20 in the population: for a population of 100,000 agents, decreased to ~ 95445 agents after removal of agents aged < 18, under $\zeta \sim 0.95$, we expect ~ 4775 agents afflicted by the chronic condition. In the next setup, the E(Y) = 0.0505

$$Y = \begin{cases} 0, X_0 + 45(X_1) + 90(X_2) + \varepsilon < 0 \\ 1, X_0 + 45(X_1) + 90(X_2) + \varepsilon > 0 \end{cases} \quad [5]$$

$$\varepsilon \sim Gaussian(\mu_\varepsilon = -295, \sigma_\varepsilon = 90)$$

σ_ε was arbitrarily set such that $\sigma \sim 6(\sigma_{X_0})$.

Issue 2: How can we simulate recruitment in the hybrid sampling?

There are two different processes to simulate a network among individuals of our population. The first is to actually employ a software that connects all the members of the virtual population in a network. The ‘primary’ sample is randomly drawn, then agents in the primary sample recruits 1 stage of secondary participants among their links. Through this method, both homophily clusters and preferential recruitment are simulated. Unfortunately, this method is unpractical for a high-sized N of population as it requires intense computational resources.

Another approach to simulate hybrid sampling is less computationally intensive:

1. the primary sample ***I*** is drawn randomly and removed from the ***N*** population. Then two new variables are assigned to any *i* agent that is an element of ***I***[6]:
   - Links:

     $$Z_1 \sim Integer\left(ChiSquared\left(df_{Z_1} = 150.5\right)\right) \quad [6]$$
   - Recruitment:

     $$Z_2 \sim Integer\left(ChiSquared\left(df_{Z_2} = 5.5\right)\right)^{7} \quad [7]$$
2. every *i* agent in the first sample randomly draws z_1 other agents j_i from the ***N*** population; j_i agents are not removed from ***N*** – ***I*** but becomes elements of $\boldsymbol{J}_\text{i}$
3. for each *i* with a $z_2 > 0$ and all its j_i, a $d_{(\text{i,j})}$ value is assigned. $d_{(\text{i,j})}$ measures statistical distance between *i* and *j* expressed by the value of a structural equation

$$D_{(\text{i,j})}(X_0, X_2) : d_{(\text{i,j})} = \left|x_{0_i} - x_{0_j}\right| + \varepsilon_d \quad [8]$$

$$\varepsilon_d \sim \left|\left(Gaussian(\mu_\varepsilon = 0, \sigma_\varepsilon \sim 0)\right) - \left(Gaussian(\mu_\varepsilon = 0, \sigma_\varepsilon \sim 0)\right)\right| \simeq 0^{8}.$$

For each *i*, z_2 elements of $\boldsymbol{J}_\text{i}$ are drawn with a probability equal to:

$$1 - \frac{d_{(i,j)}}{\Sigma d_{j_i}} \quad [9]$$

which is the function of preferential recruitment. In order to reduce endogenous bias introduced by how the structural equation of distance $D_{(\text{i,j})}$ is modelled we perform a re-sample: a sample equal in size of ***I*** is randomly re-sampled into ***I’*** from the ∪(***J)*** union of all the *j* agents drawn from each *i*. The union of ***I*** and ***I’*** is the actual ‘hybrid sample’.

In this model only age has an impact on statistical distance. Impact of sex and other traits was already a topic in Coleman (1958), but since ***E***(X_1) and ***E***(X_2) are equal per side, any non-complex interaction among endogenous variables in $D_{(\text{i,j})}$ would result in a more ‘randomised’ estimates in the final sample.

3 Results and conclusions

The simulation and all the procedures were performed with software *R*[9].

[6] This design reflects “layers of friendships” mentioned in Mac Carron, Kaski, and Dunbar (2016). The parameters are then chosen only as standards for the recruitment process.

[7] *Integer* is a function that subtracts to a value its mantissa. This operation was made to force integer values on the random variables. 0.5 was added to the parameters in order to compensate the loss of the mantissa. A different approach is to model the random variables from Poisson’s function, or from an exponential one.

[8] While in this model the distance error ε_d representing exogenous factors of $D_{(\text{i,j})}$ is assumed to be equal to 0 by forcing $\sigma_\varepsilon \sim 0$, we kept it in the equation the normal model that in our opinion best fit a random error in distance for future developments.

[9] Packages: base, survival, dplyr, rlist.

The population of N ~ 95445 was simulated only once. The random sample was drawn with a size equal to 1056 agents, which is the minimum size of representative random sample for confidence interval of 3% and a CL equal to 95%. Hybrid sample was drawn from the same population but with a size of ***I*** equal to 528, so the size of the final sample ***I*** ∪ ***I'*** is again 1056.

Both the design were run and recorded 300 times. We noticed that the average amount of ***J*** ~ 2380. Hence, we decided to perform a second hybrid sampling with ***I*** = 264 and ***I'*** = 792. Statistics are reported in table 1 and table 2.

Table 1. Population statistics.

	N	***X_0***		***X_1***		***X_2***		***Y***	
		μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
Simulated Population	95443	46.12	284.86	0.5	0.25	0.5	0.25	0.05	0.047

Table 2. Bootstrap statistics (n=300) of samples in random and hybrid sampling design.

	n	***X_0***		***X_1***		***X_2***		***Y***	
		Mean	***Var***	***Mean***	***Var***	***Mean***	***Var***	***Mean***	***Var***
Random Sampling	300	46.15	0.22	0.49	~0	0.49	~0	0.049	~0
Hybrid Sampling (stage 0 = 528)	300	45.95	0.21	0.5	~0	0.49	~0	0.05	~0
Hybrid Sampling (stage 0 = 264)	300	47.5	0.25	0.5	~0	0.49	~0	0.051	~0

We checked if the difference between the mean of population (46.12) and the bootstrap means of all the variables in the random and hybrid samplings (see Table 2) are significant with *p*-value < 0.05. We found:

- no significant difference in all variables in random design, as expected
- no significant differences in variables of hybrid design (stage 0 = 528)
- significant difference in X_0 and Y in hybrid design (stage 0 = 264).

For this reason, we exclude feasibility of hybrid sampling (stage 0 = 264). On the basis of the results from our experiment, we can assert that exists at least a setup of very simple parameters under which hybrid model can be employed instead of random sampling design.

We think the variances in bootstrap statistics are abysmal because complexity in population data is low and recruitment processes show a low variance (i.e. Chi Square functions are leptokurtic).

4 Future developments

We propose the following topics as future developments of the present study:

1. Complexity in the population's equations: in empirical demographic data, age is not independent from sex. Behaviours are complex: some may be dependent from age and sex while others may not. For future applicative applications, the target behaviour may be modelled as a structural equation. A dummy model may result simplistic, so the output may be rescaled into a multipoint scale, instead. Chronic conditions may be results of non-linear interactions among factors. The final model could be driven from a scientific theory. For internal validity of the model, the relevant parameter is the skewness of the variables in the population: if the variables are not symmetrical, even in a population with zero homophily, where ***J***$_i$ are randomly sampled, and if preferential recruitment is modelled such that agents in the ***I*** subsample 'prefer' to draw other agents that are not statistically distant from them, then we expect an increase in estimates' variance.
2. Complexity in the preferential recruitment: we suppose this point is both the most controversial and the most impactful on biases in asymptotical estimates. We already stated that in order to simulate homophily, we would need to simulate a network. We think a practical compromise could

be to import a network dataset and then test sampling through it, although this incurs in issue mentioned in section 2. We admit we don't know how people recruit other respondents into survey tool, neither we feel like we can generalise too much on this issue. Our general intuitions for future developments are: (i) Z_2 should be platykurtic and with a fat right tail, i.e. an exponential distribution. The fatter the right tail, the more the model is stressed; (ii) the weight in likelihood of recruitment should be positively correlated to a statistical distance.

3. Multi-stage and stage-free models: in order to stress more a reduction in size of 'stage 0' ***I***, more stages of recruitment can be added, so that, i.e. a two-staged hybrid sampling would sample the union of ***I, I', I''***, etc., where the latter stage is recursively not randomly drawn among 'friends' of the precedent stage. If a development of the model simulates a network population with complex variables, a more realistic approach may be free of stages: in other terms, every agent can recruit another, until the union of all sampled agents is equal to target *n*.
4. To evaluate with a regressive model the fitness of bootstrap statistics: in the present study, the validity of the sampling design was tested through significance of difference between bootstrap statistics and population parameters. A different evaluative approach is to estimate fitness of bootstrap statistics in a regressive model for each Y, already known as outcome of structural equation.

References

1. Alfons, A., Kraft, S., Templ, M., & Filzmoser, P. (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. Stat Method Appl, 20(3), 383–407.
2. Atkinson, R. & Flint, J. (2001). Accessing Hidden and Hard-to-Reach Populations: Snowball Research Strategies. Social Research Update, 33.
3. Bryc, W. (1995). The Normal Distribution: Characterizations with Applications. Springer-Verlag.
4. Burgard, J. P., Kolb, J.-P., Merkle, H., & Münnich, R. (2017). Synthetic data for open and reproducible methodological research in social sciences and official statistics. AStA Wirtschafts- Und Sozialstatistisches Archiv, 11(3-4), 233–244.
5. Coleman, J. (1958). Relational Analysis: The Study of Social Organizations with Survey Methods. Hum Organ, 17(4), 28–36.
6. Craig, B. M., Hays, R. D., Pickard, S. A., Cella, D., Revicki, D. A., & Reeve, B. B. (2013). Comparison of US Panel Vendors for Online Surveys, J Med Inter Res, 15(11).
7. Crawford, F. W., Aronow, P. M., Zeng, L., & Li, J. (2017). Identification of Homophily and Preferential Recruitment in Respondent-Driven Sampling. Am J Epidemiol, 187(1), 153–160.
8. Erickson, B. H. (1979). Some Problems of Inference from Chain Data. Sociol Methodol, 10, 276.
9. Etter, J.F., & Perneger, T. V. (2000). Snowball sampling by mail: application to a survey of smokers in the general population. Int J Epidemiol, 29(1), 43-48.
10. Gil, M. Á., Montenegro, M., González-Rodríguez, G., Colubi, A., & Rosa Casals, M. (2006). Bootstrap approach to the multi-sample test of means with imprecise data. Comput Stat Data An, 51(1), 148–162.
11. Gobet, E. (2016). Monte-Carlo Methods and Stochastic Processes. New York: Chapman and Hall/CRC.
12. Goodman, L.A. (1961). Snowball sampling. Annals of Mathematical Statistics. 32(1), 148–170.
13. Goodman, L.A. (2011). Comment: On Respondent-Driven Sampling and Snowball Sampling in Hard-to-Reach Populations and Snowball Sampling Not in Hard-to-Reach Populations. Sociol Methodol, 41(1), 347–353.
14. Kline, P. (2015). A Handbook of Test Construction (Psychology Revivals). London: Routledge.
15. Mac Carron, P., Kaski, K., & Dunbar, R. (2016). Calling Dunbar's numbers. Soc Netw, 47, 151–155.
16. Newman M (2010). Networks: An Introduction. New York, NY: Oxford University Press.
17. Ricklefs, R. &Scheuerlein, A. (2002). Biological implications of the Weibull and Gompertz models of aging. J Gerontol A-Biol, 57(2), 69-76.
18. Snijders, T. A. B. (1992). Estimation on the Basis of Snowball Samples: How to Weight? BSM, 36(1), 59–70.
19. Siddiqui, N. A., Rabidas, V. N., Sinha S. K., Verma R. B., Pandey, K. P., Singh, V.P., Ranjan, A., Topno, R. K., Lal, C. S., Kumar, V., Sahoo, G.C., Sridhar, S., Pandey, A., Das, P. (2016) Snowball Vs. House-to-House Technique for Measuring Annual Incidence of Kala-azar in the Higher Endemic Blocks of Bihar, India: A Comparison. PLoS Neglected Tropical Diseases, 10(9).