SAE2021 BIG4small Book of short papers

Editors: Monica Pratesi, Roberta Siciliano, Gaia Bertarelli, Antonio D'Ambrosio

Edizioni Zaccaria, Napoli - Italy
December 2021
ISBN 9788899594138

Preface

Planning and evaluation of government programmes usually requires access to a huge amount of data concerning national, sub-national and supranational socio-economic, environment and health related statistics. There is, however, a growing need for statistics relating to much smaller geographical areas, where data are too sparse to support the sort of standard estimation methods typically employed at national level. These *small area* official statistics are routinely used for a variety of purposes, including assessing economic well-being of a nation, making public policies, and allocating funds in various government programmes. With the rapid development of survey methodology, different governmental agencies are now exploring ways of combining national survey data with a variety of structured and unstructured data, including administrative, census records and Big Data to produce reliable small area statistics. The field of small area estimation research is quickly expanding to meet this demand, and is constantly tackling practical problems that are theoretically challenging.

The **SAE2021: Conference on Big Data for Small Area Estimation - BIG4small** (SAE2021 - `www.sae2021.org`) - a Satellite Meeting of the International Statistical Institute 63rd World Statistics - collected talks that tried to address these issues from a methodological and/or applicative point of view, together with traditional topics in estimating for small areas.

The conference program has included 4 plenary sessions, 15 invited sessions, 8 solicited sessions and the award for Outstanding Contribution to Small Area Estimation 2020 and 2021 ceremony, assigned to Partha Lahiri and Wayne A. Fuller, respectively. The conference committee had registered 73 accepted submissions, including 49 to be presented in plenary and invited sessions, and 24 spontaneously submitted for oral solicited sessions. The scientific program can be found at the end of this preface.

For more than a year, the Covid-19 pandemic has hit our most consolidated habits with serious challenges on social and economic system. Implementation of the guidelines for social distancing has led to the shifting of most of the research activities remotely. After very careful consideration, concerning the health of all conference participants and the restricted mobility of the staff of many universities and research centres, the Advisory Board and the Local Organizing Committees decided to schedule the SAE 2021 in remote from the 20th to the 24th of September 2021. The Conference was streamed through the Zoom platform provided by the University of Naples Federico II. The organizers met at the University of Naples Federico II during the management of the conference.

This volume gathers part of the peer-reviewed papers submitted to the BIG4small SAE 2021 conference. The volume covers a wide variety of subjects ranging from methodological and theoretical contributions, to applied works and case studies, giving an excellent overview of the interests of the international researchers who work on SAE topic.

Of course, both the Conference and this volume would not be possible without the collaboration of the members of the Programme Committee and the members of the University of Naples Federico II and of the University of Pisa. Members of these two institutions took part actively in the Local Organizing Committee. The conference also received support from sponsors, namely the International Association of Survey Statisticians (IASS - `http://isi-iass.org/home/`) and the Department of Economics and Statistics (DISES - `http://www.dises.unina.it/`), University of Naples Federico II. To all of them, our thanks.

Our thanks also go to the keynote lecturers, all contributors for having submitted their work to the conference, the members of the Advisory Board, the Programme Committee and the extra reviewers for their efforts in this difficult period.

Monica Pratesi
University of Pisa, Italy (Chair of the Program Committee)

Roberta Siciliano
University of Naples Federico II, Italy (Chair of the Program Committee)

Antonio D'Ambrosio
University of Naples Federico II, Italy (Chair of the Organizing Committee)

Gaia Bertarelli
Sant'Anna School of Advanced Studies, Pisa, Italy.

Keynote Lecturers: Graham Kalton, J. Sunil Rao, Andrea Saltelli.

Advisory Board: Partha Lahiri, J.N.K. Rao, Malay Ghosh, Carlo N. Lauro, Giuseppe Longo, Danny Pfeffermann.

Program committee: Monica Pratesi (Chair), Roberta Siciliano (Chair), Giovanni Acampora, Massimo Aria, Gaia Bertarelli, Jan A. van den Brakel, Dario Buono, Sanjay Chaudhury, Marcello Chiodi, Cinzia Cirillo, Antonio D'Ambrosio, Sun Dongchu, Stefano Falorsi, Julie Gershunskaya, Ornella Giambalvo, Jiming Jiang, Domingo Morales, Ralf Muennich, Valentin Patilea, Alessandra Petrucci, Isabel Molina Peralta, Maria Giovanna Ranalli, Nicola Salvati, Germana Scepi, Timo Schmid, Natalie Shlomo, Marcyn Szymkowiak, Nikos Tzavidis, Linda J. Young, Rosanna Verde.

Local committee: Antonio D'Ambrosio (Chair), Germana Scepi (Chair), Enrico Cafaro, Carmela Cappelli, Carmela Iorio, Giuseppe Pandolfo, Alfonso Piscitelli, Rosaria Romano, Maria Spano, Michele Staiano, Autilia Vitiello.

Conference program

Monday, September 20 - **Day 1**	
01:00 PM – 01:30 PM CEST Main Room 2	**Opening and welcome** *Matteo Lorito*, Rector Magnificus of University of Naples Federico II *Monica Pratesi*, President of IASS & Head of the Programme Committee *Roberta Siciliano*, Head of the Programme Committee *Antonio D'Ambrosio*, Head of the Local Committee
01:30 PM - 02:30 PM CEST Main Room 1	**Plenary session 1** Chair: *Roberta Siciliano* - Discussant: *Monica Pratesi* From Survey Sampling to Small Area Estimation - *Graham Kalton*
02:30 PM – 02:45 PM CEST	Break
02:45 PM – 04:15 PM CEST Main Room 2	**Invited Sessions 1: Bayesian Statistics for Small Area Estimation** – Chair: *Partha Lahiri* Shrinkage Estimation with Singular Priors and an Application to Small Area Estimation - *Malay Ghosh* Small Area Estimation of Vaccination Coverage Using Non-Survey Sources - *rivellore Raghunathan* Pseudo Bayesian Estimation of one-way anova model in complex survey - *Terrance D. Savitsky*
04:15 PM – 04:30 PM CEST	Break
04:30 PM – 05:30 PM CEST Room 1	**Solicited Session 1** – Chair: *Gaia Bertarelli* Small Area Estimation of Monetary Poverty in Mexico using Satellite Imagery and Machine Learning- David Newhouse Incidence of poverty in Costa Rica: small area estimates under a Structure Preserving Estimation (SPREE) approach - Alejandra Arias-Salazar Leave No One Behind: SDG Monitoring using Small Area Estimation in Latin America - Andres Gutierrez
04:30 PM – 05:30 PM CEST Room 2	**Solicited Session 2** – Chair: *Carmela Cappelli* Skew-Normal CAR models for small domain estimation in the Brazilian Annual Service Sector Survey - *Andre Felipe Azevedo Neves* Estimation of the Employment Rate by Municipality in Mexico. Interpreting Results from an SAE - *model - Enrique de Alba* Small Area Estimates of Labor Force Statistics in Urban Mexico using Geospatial Data - *Joshua D. Merfeld*
05:30 PM – 07:00 PM CEST Main Room 1	**Invited Sessions 2: Time series methods for Small Area Estimation** – Chair: *Jan Van Den Brakel* Multilevel time-series models for estimation at different frequencies and regional levels - *H.J. Boonstra* Multilevel time series modeling of mobility trends in the Netherlands for small domains - *S. Das* Some Issues in Seasonal Adjustment of Time Series from Repeated Sample Surveys - *W.R. Bell*

Tuesday, September 21- **Day 2**	
10:30 AM – 11:30 AM CEST Room 1	**Solicited Session 3** – Chair: *Alfonso Piscitelli* Flexible Small Area Estimation of Theil Index using Mixture of Beta - *Silvia De Nicoló* On properties of MSE estimators of the EBLUP for some class of Linear Mixed Models in small area estimation - *Malgorzata Krzciuk* Inference on quantiles in small area based on estimates of the distribution function - *Tomasz Stachurski*
10:30 AM - 11:30 AM CEST Room 2	**Solicited Session 4** – Chair: *Giuseppe Pandolfo* Using Random Forests in SAE - *Patrick Krennmair* The comparison of different machine learning methods in small area prediction problems - *Adam Chwila* Statistical data integration as an extension of small area estimation for employee compensation - *Andreea Luisa Erciulescu*
11:30 AM – 01:00 PM CEST Main Room 1	**Invited Sessions 3: Young Researchers' contribution to Small Area Estimation** - Chair: *Gaia Bertarelli* - Discussant: *Angelo Moretti* Variable and Transformation Selection for Linear Mixed Models with Application to Small Area Estimation - *Yeonjoo Lee* Time stable empirical best predictors under a unit-level model - *Maria Guadarrama* Controlling the bias for M-quantile estimators for small area - *Francesco Schirripa Spagnolo*
01:00 PM – 01:30 PM CEST	Break
01:30 PM – 03:00 PM CEST Main Room 2	**Invited Sessions 4: Issues and opportunities from record linkage and data integration in Small Area Estimation** - Chair: *Silvia Polettini* Record linkage, measurement error and unit level small area estimation: a Bayesian approach - *Serena Arima* Small area estimation in a linkage errors framework: area-level vs unit-level models - *Loredana Di Consiglio* Error-in-covariates in small area estimation and a generalized Fay-Herriot Model - *Gauri Datta*
03:00 PM – 03:15 PM CEST	Break
03:15 PM – 04:15 PM CEST Main Room 1	**Plenary session 2** Chair: *Monica Pratesi* - Discussant: *Roberta Siciliano* A Tour of Classified Mixed Model Predictions and Projections - *J. Sunil Rao*
04:15 PM - 05:45 PM CEST Main Room 2	**Invited Sessions 5: Selected Challenges is Small Area Estimation** - Chair: *Ralf Münnich* Empirical best prediction of bivariate nonlinear small area indicators - *Domingo Morales* On "qape" R package for measuring accuracy of small area predictors - *Thomas Zadlo* Regularized Small Area Estimation: A Framework for Robust Estimates in the Presence of Unknown Covariate Measurement Errors - *Joscha Krause*
05:45 PM - 07:15 PM CEST Main Room 1	**Invited Sessions 6: Disaggregated data and indicators from Big data sources** - Chair: *Rosanna Verde* - Discussant: *Monica Pratesi* Social networks data and small area estimation: a tentative solution to overcome selection bias - *Elena Siletti* Small area poverty indicators adjusted using local price indexes - *Caterina Giusti* Small area estimation via Heteroskedastic Geographically Weighted Regression for functional data - *Elvira Romano*

Wednesday, September 22 - **Day 3**	
10:30 AM - 11:30 AM CEST Room 1	**Solicited Session 5** - Chair: *Carmela Iorio* Discovering Dynamics in Land Systems using Time Series Analysis and Non-linear Dynamical Methods - *Richard Aspinall* Small Area Estimation of Growing Stock Volume with Fay-Herriot area-level model - *Aristeidis Georgakis*
10:30 AM - 11:30 AM CEST Room 2	**Solicited Session 6** - Chair: *Francesco Schirripa Spagnolo* On benchmaking small area estimators when the model is misspecified - *Renato Salvatore* Hierarchical Bayesian Spatial Small Area Model for Binary Data Under Spatial Misalignment - *Kindie Fentahun Muchie* The inverse sampling method in the Big Data Era - *Daniele Cuntrera*
11:30 AM - 01:00 PM CEST Main Room 1	**Invited Sessions 7: Small area estimation for latent variables and complex indicators** - Chair: *Maria Giovanna Ranalli* - Discussant: *Gaia Bertarelli* Estimating small area latent social integration of second-generation students in Italy - *Francesco Giovinazzi* Unit level models on the log-scale: a new Bayesian proposal for poverty mapping - *Aldo Gardini* Multivariate small area estimation methods for multidimensional latent wellbeing indicators - *Angelo Moretti*
01:00 PM - 01:30 PM CEST	Break
01:30 PM - 03:00 PM CEST Main Room 2	**Plenary Session: Hukum Chandra Memory** - Chair: *Raymond Chambers* Hukum Chandra: A Tribute - *J.N.K. Rao* Contribution of Dr. Hukum Chandra in the Theory and Applications of Survey Sampling - *Priyanka Anjoy* Hukum's work on outlier robust SAE and Non-spatial stationarity models for SAE - *Nicola Salvati & Nikos Tzavidis* NSAE: An R Package for Small Area Estimation under Spatial Nonstationarity - *Saurav Guha*
03:00 PM - 03:15 PM CEST	Break
03:15 PM - 04:45 PM CEST Main Room 1	**Invited Sessions 8: Small Area Estimation in Official Statistics** - Chairs: *Nicola Salvati & Francesco Schirripa Spagnolo* Small area estimates of labour market status using multinomial expectile regression - *Enrico Fabrizi* Robust small area estimation in business surveys - *Chiara Bocci* Causal inferences for official statistics - *Setareh Ranjbar*
04:45 PM - 05:00 PM CEST	Break
05:00 PM - 06:30 PM CEST Main Room 2	**Invited Sessions 9: Recent Advances in Model Selection and Diagnostics for Small Area Estimation** - Chair and Discussant: *Jiming Jiang* Selection of auxiliary variables for two-fold subarea-level linking models in small area estimation: A simple method - *J.N.K. Rao* A Robust Goodness-of-fit Test for Small Area Estimation - *Mahmoud Torabi* Recent Advances in Measures of Uncertainty in Post Model Selection Small Area Estimation - *Thuan Nguyen*

Thursday, September 23 - **Day 4**	
10:30 AM - 12:00 PM CEST Main Room 1	**Invited Sessions 10: Small Area Estimation for Permanent population Census and Social Surveys: new applications and methods** - Chair: *Stefano Falorsi* MIND, an R Package for multivariate and multiple random effect small area estimation - *Andrea Fasulo* SAE estimation under coherence for different overlapping areas. An application for the estimation of employment and unemployment from LFS for cities and FUAs - *Silvia Loriga* Defining the sample designs for small area estimation - *Piero Falorsi*
12:00 PM - 01:30 PM CEST Main Room 2	**Invited Sessions 11: Small Area estimation: new developments and applications** - Chair: *Carmela Cappelli* A Hierarchical Bayesian Approach for Addressing Multiple Objectives in Poverty Research for Small Areas - *Stefano Marchetti* Best Prediction of Missing Area-Level Direct Estimates via Multivariate Modelling - *Anna-Lena Wölwer* Small area estimation via multivariate generalized linear mixed effects models - *Emilia Rocco*
01:30 PM - 02:30 PM CEST	Break
02:30 PM - 03:30 PM CEST Main Room 2	**Invited Sessions 12: Some novel developments in small area estimation** - Chair: *Sanjay Chaudhuri* Robust, high-dimensional data linkage for small area statistics - *Snigdhansu Chatterjee* Covariance based Moment Equations for Improved Variance Component Estimation - *Sanjay Chaudhuri*
03:30 PM - 03:45 PM CEST	Break
03:45 PM - 04:45 PM CEST Main Room 1	**Plenary Session 3** Chair: *Gaia Bertarelli* - Discussant: *Antonio D'Ambrosio* Ethics of Quantification - *Andrea Saltelli*
04:45 PM - 05:00 PM CEST	Break
05:00 PM - 06:00 PM CEST Room 1	**Solicited Session 7** - Chair: *Francesco Schirripa Spagnolo* Estimation of life expectancy in small areas using big data from the municipal registry - *Carlo Cusatelli* Reliable event rates for disease mapping - *Harrison Quick*
05:00 PM - 06:00 PM CEST Room 2	**Solicited Session 8** - Chair: *Carmela Cappelli* Design-based small area estimation: an application to the DHS surveys - *Ruilin Ren* Design-based composite estimation of small proportions in small domains - *Andrius Ciginas* Challenges and lessons learned in using small area estimation for official statistics "how could we help?" - *Haoyi Chen*

Friday, September 24 - **Day 5**	
10:30 AM - 12:00 PM CEST Main Room 2	**Invited Sessions 13: Data Science Methodology Transfer: Big to Small** - Chair: *Giuseppe Longo* - Discussant: *Giuseppe Pandolfo* Bias versus statistical errors in Big data information systems - *Edwin A. Valentijn* Using proper scoring rules to derive well calibrate photometric redshift models - *Kai Polsterer* Error Mitigation in Quantum Measurement through Fuzzy C-Means Clustering - *Autilia Vitiello*
12:00 PM - 12:15 PM CEST	Break
12:15 PM - 01:45 PM CEST Main Room 1	**Invited Sessions 14: Latent variables in small are models: theoretical and applied issues** - Chair and Discussant: *Serena Arima* Empirical Best Prediction for Small Area Estimation of categorical variables using Finite Mixtures of Multinomial Logistic Models - *Maria Giovanna Ranalli* Small area models with uncertainty on measurement error in covariates - *Silvia Polettini* Bayesian model selection for log-linear latent class models - *Davide Di Cecco*
01:45 PM - 02:15 PM CEST	Break
02:15 PM - 03:45 PM CEST Main Room 2	**Invited Sessions 15: Inference under informative sampling** - Chair: *Danny Pfeffermann* Informative or ignorable selection process: a review - *Daniel Bonnery* Spatial processes and endogenous spatial selection, estimation and prediction - *Francesco Pantalone* An Approximate Best Prediction Approach to Small Area Estimation for Sheet and Rill Erosion under Informative Sampling - *Emily Berg*
03:45 PM - 04:45 PM CEST Main Room 1	**Award of Outstanding Contribution to Small Area Estimation**
04:45 PM - 05:00 PM CEST Main Room 1	**Closing Ceremony**

Contents

Discovering Dynamics in Land Systems using Time Series Analysis and Non-linear Dynamical Methods

Richard Aspinall [a]*, Michele Staiano [b]**, Diane Pearson [c]*** and Alfonso Piscitelli [d]****

*Honorary Research Fellow James Hutton Institute, Aberdeen, UK
**Department of Industrial Engineering, University of Naples Federico II, IT
***School of Agriculture and Environment, College of Sciences, Massey University, NZ
****Dep. of Agricultural Sciences, University of Naples Federico II, IT

Abstract

We use time series analysis, including methods from for non-linear dynamical systems, to separate different types of dynamics encapsulated within a historical record of land system states for farming in Scotland over the period from 1867 to 2020. Our results characterize dynamics from internal feedbacks and coupling of farming as a system at the national scale, reveal some system characteristics and behaviours associated with the dynamical evolution of farming as a system, and identify some regime shifts over the full 154-year timespan of the Scottish agricultural census. Specifically, the results reveal i) consequences of several exogenous factors as events that had an impact on system states, ii) show that arable and pastoral farming, at a national scale, are dynamically related over a range of timescales and coupled to global trends, and iii) that throughout much of the timespan of the study the system has maintained a pattern of changes consistent with endogenous systems-level feedbacks between sectors that act to dampen the impacts of exogenous factors. Changes in system dynamics over the timespan are also associated with policy changes that altered the interaction of arable and pastoral farming. Analysis of data for counties within NE Scotland show similar trends and suggest that

[a] rjaspinall10@gmail.com
[b] mstaiano@unina.it
[c] d.pearson@massey.ac.nz
[d] alfonso.piscitelli@unina.it

the patterns discovered in national aggregate data are informative about farming in smaller areas.

Keywords— land system dynamics, time series analysis, non-linear dynamical systems, Scotland

1 Introduction

Dynamics in land systems are associated with process–response (cause and effect) relationships, endogenous behaviours resulting from system–level interactions, and adaptation to exogenous factors. Discovering these dynamics for different land uses, geographical contexts, and historical periods is a core activity of land systems science.

In this paper, we use time series analysis, including methods for non-linear dynamical systems, to separate dynamics encapsulated within a historical record of land system states for farming in Scotland over the period from 1867 to 2020, and for smaller geographic areas within Scotland from 1892 to 1975. The data are from the annual series of agricultural census, that provides a summary of farming in Scotland. Although these data record the annual state of farming, they are seldom analysed for trends beyond short-term changes or for information beside the status of different components of the funds of land use, livestock numbers and productivity that define the national account of farming. We use the data at national and county scales, with a simple systems model of farming land use as a coupled human-environment system, to elucidate information on dynamics of farming and its evolution over time.

2 Data

The data used are time series of annual records describing the farming system in Scotland from 1867 to 2020 and from 1892 to 1975 for five counties in NE Scotland. Data describe area planted in cereals and the number of sheep, both compiled from the Annual Agricultural (June) Censuses of Scotland which have been published over the last 154 years.

The total planted area in cereals is used to represent the arable sector. The number of sheep are used to represent the pastoral. These national and county-level data are collated from individual farm-level survey returns; clearly the participants in the survey have changed over time, as have the nature of the farming systems being recorded. Price of barley for each year is from multiple sources, including annual reports on Agricultural Statistics (1912-1978), Economic Reports on Scottish Agriculture (1980-2020. All prices data are converted to pounds sterling from a variety of source prices (viz. Scots and English pounds, shillings and pence (£/s/d), GB Pounds after decimalisation in 1971) in use at the time of original data collection.

3 Methods

We analyse dynamics using two set of methods: i) time series analysis, including lag plots and decomposition of time series into long-, medium-, and short-term components; and ii) Recurrence plots (RP) (Eckmann et al., 1998; Marwan et al., 2007) and Recurrence Quantification Analysis (RQA) (Webber and Marwan, 2015) from analysis of non-linear dynamical systems. Use of different methods in combination helps to assess the coherence and reliability of the data from the time series.

4 Results

The time series analysis for barley price, cereal area and sheep numbers for Scotland are shown in Figure 1; the recurrence plots, RP, and recurrence quantification analysis, RQA, for these data are shown in Figure 2. The long-term trend for barley price is exponential growth (Figure 1a). Deviations from the long-term trend are modelled with a smoothing spline, revealing 4 main cycles over the 154-year period (Figure 1b). The long-term trend in cereal area is an annual decline in area planted of 0.14%, accumulating to a total of about 19% over the 154-year period (Figure 1d), with four cycles superimposed on the long-term trend (Figure 1e). The long- term trend for sheep is of increasing numbers (Figure 1g), with five medium term cycles (Figure 1h). The RP for barley price (Figure 2a) shows clear evidence of regime shifts, with areas of black along the diagonal of the plot and no recurrence points outside those boxes. The regime shift in the 1970s is clear in the time series plot (Figure 1a), but the RP shows there was a further shift starting in the 1930s, and another in the 1950s lasting into the 1960s. The white bars coinciding with the world wars indicate extreme variability in barley price; high variability since 1973 is also revealed in the absence of recurrence points. The RP for cereals (Figure 2b) shows long term cycles, with variability during the two world wars; the RP also shows increased cycles in the period since 1973. The RP for sheep (Figure 2c) shows clear evidence of cycles in the number of sheep, with regular pattern of recurrences spaces about 30 years apart, three cycles being evident since 1950; numbers were more stable in the latter part of the 19th century. Figure 3 shows the medium-term cycles for cereal area and sheep numbers for counties in NE Scotland, as well as for NE Scotland and Scotland, as well as barley price for Scotland, from 1892-1975. This plot shows the close interdependence and synchronisation between the variables in the medium-term, making evident their associations: sheep numbers are negatively correlated with cereal area and barley price, and barley price and cereal area are positively correlated. As barley (and other cereal) prices increase cereal area increases and sheep numbers fall; as barley price falls the area of cereals also falls and sheep numbers increase. From 1973-2020, when the UK was within the EEC/EU Common Agricultural Policy, the associations between the three variables break down as cereal production and sheep farming, representing the arable and pastoral systems, become somehow uncouple.

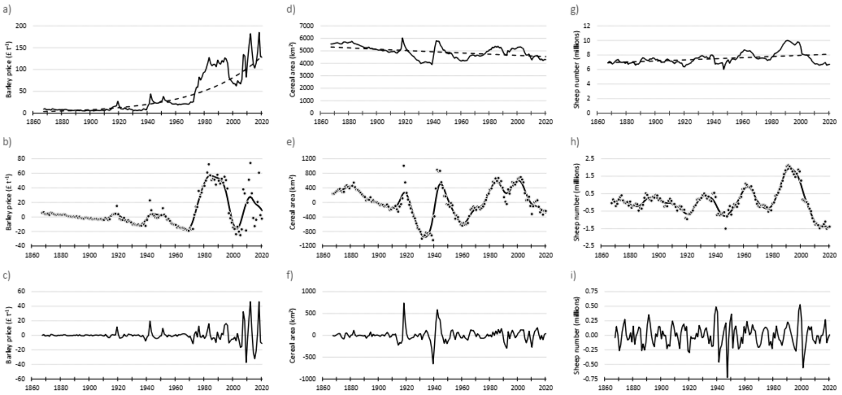


Figure 1: Time series results for barley price, cereal area and sheep numbers in Scotland, 1867-2020

5 Discussion

Trends and cycles over different timespans and timescales identified within the data using time series analysis, as well as RP and RQA, characterise long-, medium-, and short-term dynamics of cereal and sheep farming and cereal prices. Irregular cycles are evident in each of barley price, cereal area, and sheep numbers, the cycles being synchronised with each other but with phase shifts. The period of these cycles is between 15 and 40 years. The long-term trends and patterns of cycles, as well as the year-to-year variability superimposed on the long- and medium-term trends, for farming, reveal the multiscale nature of temporal variation in changes to farming systems. Cycles in the data for smaller areas within NE Scotland (from 1892-1975) are consistent with the national and regional pattern, suggesting, for NE Scotland in this instance, that the national pattern is informative about regional trends. The RP and RQA also help to identify regime shifts. In particular, the RP (Figure 2a) and RQA (Figure 4) for Barley price shows clear evidence of regime shifts, with one regime over the period from 1867 to the late-1930s (interrupted by World War one), and two further shifts in about 1950 and 1970; since 1970 the price has been highly volatile. Regime shifts are not apparent for cereal area and sheep number.

In summary, in systems terms the analysis of the historical record of changes in cereal area and price, grassland area, and sheep numbers in Scotland reveals a complex pattern of interdependencies and coupling over time and at different scales, combining endogenous system dynamics with short-term variability associated with stochastic events, within a broader set of higher-level interdependencies and boundary conditions for the system. The long

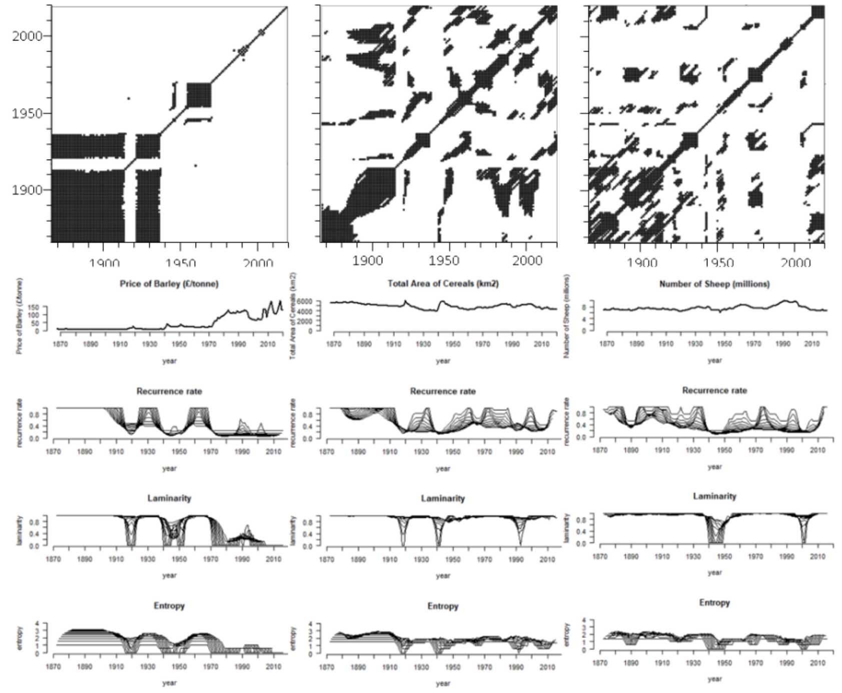


Figure 2: Recurrence plots and RQA for barley price, cereal area and sheep numbers in Scotland, 1867-2020

time-period of the study also shows that the embedded system dynamics can make farming relatively resilient to changes in policy, exogenous shocks (e.g., weather events or disease outbreaks), or regime changes and thresholds (as seen here in prices). These results, taking a long-term, whole systems perspective, and use of methods from time series analysis, reveals the evolution of land use as a dynamic and dynamical system. Moreover, exploring recurrence as a fundamental property of dynamical systems, the proposed RP-RQA analysis shows interesting potential to be applied for effectively capturing relationships (viz. associations and interdependences) between a whole system perspective and small areas.

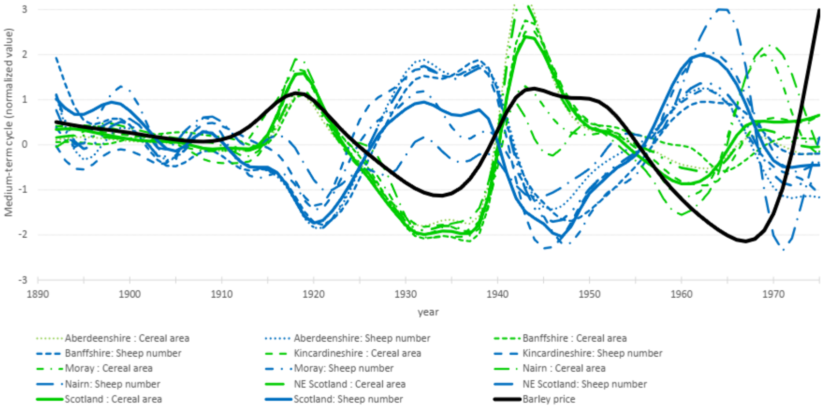


Figure 3: Medium-term cycles in cereal area and sheep numbers for counties in NE Scotland, NE Scotland and Scotland and barley price for Scotland (1892-1975)

References

Eckmann, J. P., Kamphorst, S. O., and Ruelle, D. (1987). Recurrence Plots of Dynamical Systems. *Europhysics Letters*, 4(9), 973–977.

Marwan, N., Carmen Romano, M., Thiel, M. and Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5), 237–329.

Webber, C. L., and Marwan, N. (Eds.). (2015). *Recurrence Quantification Analysis: Theory and Best Practices*. Springer, London.

Small area poverty indicators adjusted using local price indexes.

Luigi Biggeri [a]*, Caterina Giusti [b]**, Stefano Marchetti[c]** and Monica Pratesi[d]**

*University of Florence - Dagum ASESD Centre
**University of Pisa - Dagum ASESD Centre

Abstract

In this work we focus on estimating monetary poverty indicators at sub-regional level in Italy taking into account the different price levels within the country. To account for the local price levels, Spatial Price Indexes (SPIs) are computed using retail scanner data on regional and sub-regional retail volumes (units) and price for food and grocery. Specifically, a Country Product Dummy model is used, with products aggregated by province and ECOICOP-8-digit classification, for a total of 103 provinces and 102 ECOICOP-8-digit. The SPIs are used to adjust the poverty line when computing provincial poverty indicators that are estimated using area level Small Area Estimation (SAE) models, which link direct unreliable estimates to aggregated auxiliary information, often easily available. The use of SAE models is necessary to obtain reliable estimates at sub-regional level that can be used to guide policy decisions to reduce poverty.

Keywords— Poverty mapping, Spatial price indexes, Scanner data, Cross product dummy model.

1 Estimation of Spatial Consumer Price Indexes for the Italian Provinces

In this work we present a methodology to compute Spatial Price Indexes (SPIs) at sub-regional level in Italy using retail scanner data on regional and sub-regional retail volumes (units) and price for food and grocery. The data were provided by Istat/Nielsen within a research project between the Dagum Center (http://www.centrodagum.it/en/) and Istat in the framework of the H2020

[a] luigi.biggeri@unifi.it
[b] caterina.giusti@unipi.it
[c] stefano.marchetti@unipi.it
[d] monica.pratesi@unipi.it

project Makswell (www.makswell.eu) (for more details please refer to the work by Pratesi, Giusti, Marchetti, Biggeri, Bertarelli, Schirripa Spagnolo, Laureti, Benedetti, Polidoro, Di Leo, Fedeli of the Makswell "Deliverable 3.2 - Guidelines for best practices implementation for transferring methodology").

Specifically, we compute SPIs for 103 (out of 110) Italian provinces, by using the scanner data referring to the year 2018 and only to the products (barcodes or Global Trade Item Numbers - GTINs) in food and beverages categories, excluding fresh food. Usually the information on products' quantities is reported in terms of grams and milliliter, but sometimes in units; given that we needed to use comparable prices, we discarded about 17,000 quotations expressed in units. To estimate the SPIs a two-step procedure has been used, adapting the World Bank Group (2015) approach.

In the first step, we computed the average unit price at provincial level, by considering the unit value prices from the consumer side. In applying the principle of comparability, we did not follow a very tight way by considering the comparisons of the 'like to like' items (products). Instead, we applied the principle at a different level, the level products' groups, and exactly at the level of the 102 groups of the ECOICOP-8-digit classification.

Define the weighted mean price $\bar{p}_{ij}$ for ECOICOP-8-digit j and province i. Let r_{ijk} and q_{ijk} be the annual turnover and the total quantity sold[1] respectively of item k belonging to ECOICOP-8-digit j in province i. These quantities are estimated by Istat using the scanner data and the sampling weights computed according to the survey design (we refer to Deliverable 3.2 of the MAKSWELL project for further details). Let u_{ijk} be the quantity of the item ijk in terms of gr. or ml. For each item we define its annual price per gr. or ml. as

$$p_{ijk} = \frac{\frac{r_{ijk}}{q_{ijk}}}{u_{ijk}}.$$

Then, for each item we define its relative weights in term of turnover as

$$w_{ijk} = \frac{r_{ijk}}{\sum_{k=1}^{n_{ij}} r_{ijk}},$$

where n_j is the number of items in the jth ECOICOP-8-digit aggregation and the ith province. Finally, the weighted mean price is:

$$\bar{p}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} p_{ijk} w_{ijk}.$$

Therefore, $\bar{p}_{ij}$ is the weighted mean price per gr. or ml. for products in ECOICOP-8-digit j and province i.

The second step is devoted to the aggregation of 102 average level of prices to estimate the provincial SPI. Note that not all the ECOICOP-8-digit aggregates are present in all the provinces.

[1] Which are the expenditure and the quantity purchased by consumers.

To compute the SPIs at provincial level we adapt a Country Product Dummy model (Laureti and Rao, 2018). The products are aggregated by province and ECOICOP-8-digit classification, for a total of 103 provinces and 102 ECOICOP-8-digit. Note that not all the ECOICOP-8-digit aggregates are present in all the provinces. The CPD model we propose is as follows:

$$\log \bar{p}_{ij} = \alpha_0 + \alpha_i D_i + \beta_j I_j + \varepsilon_{ij}, \quad i = 1, \dots, 103 \quad j = 1, \dots, 102, \qquad (1)$$

where D_i is a vector equal 1 if the mean price is in province i and 0 otherwise, I_j is equal 1 if the mean price belongs to jth ECOICOP-8-digits and 0 otherwise. The index i is for the provinces and the index j is for the ECOICOP-8-digit. The error $\varepsilon_{ij} \sim N(0, \sigma^2)$.

To take into account the different level of the turnover between the ECOICOP-8-digit aggregates we estimate the model (1) using weighted least squares, where the weights are computed as

$$wls_{ij} = \frac{\sum_{k=1}^{n_{ij}} r_{ijk}}{\sum_{k=1}^{n_i} r_{ijk}},$$

the ratio between the total turnover of one aggregate in one province and the total turnover in the province (n_i is the number of items in the ith province).

Model (1) – as it is specified – is not identified, because the D_is vectors are a linear combination of the constant. Therefore, we impose the constraint $\alpha_1 = 0$ so that the model is identified. Once the model is estimated, from the data we obtain the estimates of the SPIs at provincial level by $\exp(\hat{\alpha}_i)$, where $\hat{\alpha}_i$ is the estimate of α_i. The coefficient α_i is the difference of fixed effects connected with the province i compared with the base province $i = 1$. To use as a reference Italy instead of area 1, the coefficients $\hat{\alpha}_i$ has been adjusted following Suits (1984).

The SPIs estimated at the province level can be used for many purposes. One of these purposes is to adjust the national poverty line at the province level, by this way relative poverty estimates take into account the different purchase power within the country. The SPIs estimated according to model (1) are based on mean prices of specific headings (ECOICOP-8-digit), therefore the adjustment of the national poverty line is not poor specific. As an alternative, the method can be easily extended to produce SPIs related to the first quintile of the distribution of the price of each specific product, assuming that poor purchase the cheaper items of the product. For example, figure 1 reports two choropleth maps of estimated SPIs based on model (1, left) and on an adjusted the model that considers the quantile 0.2 of the unit prices to obtain the estimates of spatial price indices related to the cheaper prices for each Italian provinces, which we denote as SPI($\mathrm{Q}_{0.2}$)'s (right).

The results we obtained are somehow expected. Indeed, provinces in the south of Italy show SPIs smaller than 1, while provinces in the north show values greater than 1. However, there are exceptions, provinces in the north-east Alps mountains show SPIs below 1, even if they are close, both considering the mean and the quantile 0.2 of unit prices.

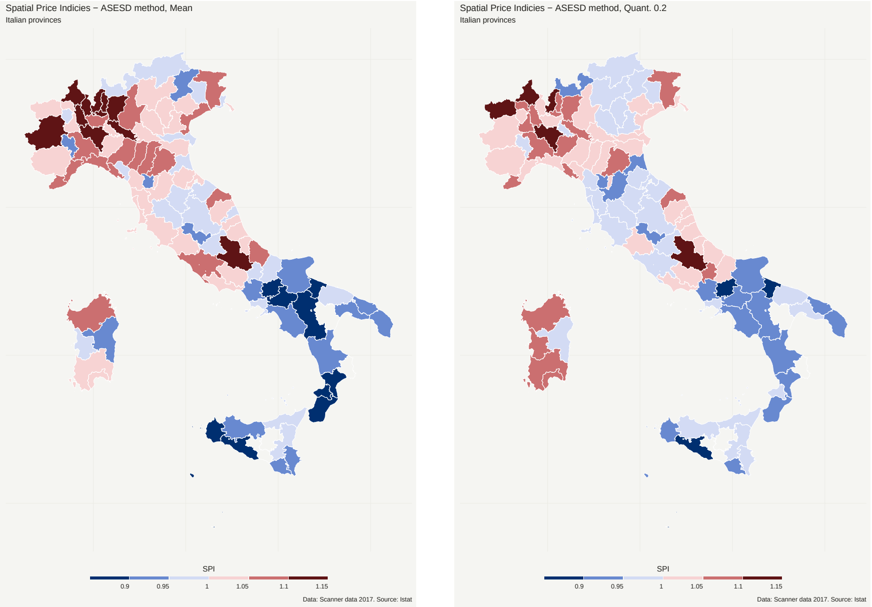

Spatial Price Indicies − ASESD method, Mean
Italian provinces
SPI
0.9
0.95
1
1.05
1.1
1.15
Data: Scanner data 2017. Source: Istat
Spatial Price Indicies − ASESD method, Quant. 0.2
Italian provinces
SPI
0.9
0.95
1
1.05
1.1
1.15
Data: Scanner data 2017. Source: Istat

Figure 1: Choropleth map of SPIs obtained using mean unit prices (left) and quantile 0.2 of unit prices (right).

2 The impact of the local cost-of-living differences on the measure of the poverty incidence

Intra-country comparisons of poverty indicators are important for many reasons. For example, when measuring the poverty incidence, the use of a national poverty line allows to establish a general scheme of how local areas (e.g. regions or provinces) compare with national standards. However, considering the same poverty line for each area implies an equity concept in which individuals with equal income are assumed to have similar wellbeing regardless of the area where they live. The use of local poverty lines allows to gauge intra-country poverty, which can be important for planning local policies. A possibile approach to compute local poverty lines is by taking into account the different price levels within the country.

In this work we estimate the Head Count Ratio – a measure of poverty incidence – at provincial level using Household Expenditure Survey (HES) data in Italy, adjusting the national poverty line using the SPI($Q_{0.2}$) values. Specifically, the national poverty line is adjusted for each province using the SPI($Q_{0.2}$) values opportunely weighted (adapting the idea in Renwick et al. (2014)):

$$nPL_i^* = nPL \times (\lambda_i SPI_i + 1 - \lambda_i) \tag{2}$$

where nPL is the national poverty line, nPL_i^* is the adjusted poverty line for province i, λ_i is the estimated share of food consumption in province i and SPI_i is the SPI($Q_{0.2}$) for province i. The quantities λ_i's are estimated from the HES 2017 as the provincial mean of the ratios between the food expenditure and the total consumption expenditure:

$$\lambda_i = \frac{1}{\sum_{j=1}^{n_i} w_{ij}} \sum_{j=1}^{n_i} \frac{p_{ij}}{t_{ij}} w_{ij}, \tag{3}$$

where n_i is the sample size in province i, w_{ij} is the survey weight of household j in area i, p_{ij} is the food expenditure of household j in area i and t_{ij} is the total consumption expenditure of household j in area i. The survey weights have been calibrated to sum to the total households at provincial level. Although the λ_i's are estimated at the provincial level – thus possibly unreliable because of small sample size – we judge the direct estimates suitable for our purpose.

Having computed the adjusted nPLs, we then calculated the corresponding direct estimates of the poverty rates. As the variability of the direct estimates was too high (approximately half of the provinces a CV greater than 30%) we estimated a Fay-Herriot (FH) model with the following auxiliary variables: the ratio between number of taxed persons over the population, and the ratios between the number of persons with *i.* income coming from salary, *ii.* income coming from pensions and *iii.* income lower than 10,000 euros per year, over the number of taxed persons. These data come from the Italian tax agency database 2017. The EBLUPs (Empirical Best Linear Unbiased Predictors) obtained with the FH model showed a gain in efficiency with respect to direct estimates. We obtained a CV smaller than 16% in 37 provinces, while half of the provinces had a CV smaller than 20%. We also computed the EBLUPs without any adjustment of the national poverty line, using the same small area model as for adjusted EBLUPs. Figure 2 reports the comparison of the two set of EBLUPs estimates: as we can see, using the SPI($Q_{0.2}$) to adjust the poverty lines, the HCRs in northern and central provinces slightly decrease.

The results obtained here suggest that the methodology can be extended to include other Spatial Price Indexes, therefore adjusting the national poverty line with other components of households' consumption expenditure. Indeed, our results suggest the products included in the scanner data represent a relevant but still limited share of the total household consumption expenditure, approximately equal to the 20%. Therefore, by including other consumption expenditure components, such as for example the expenditure for the rent, the national poverty line could be adjusted in a more complete manner.

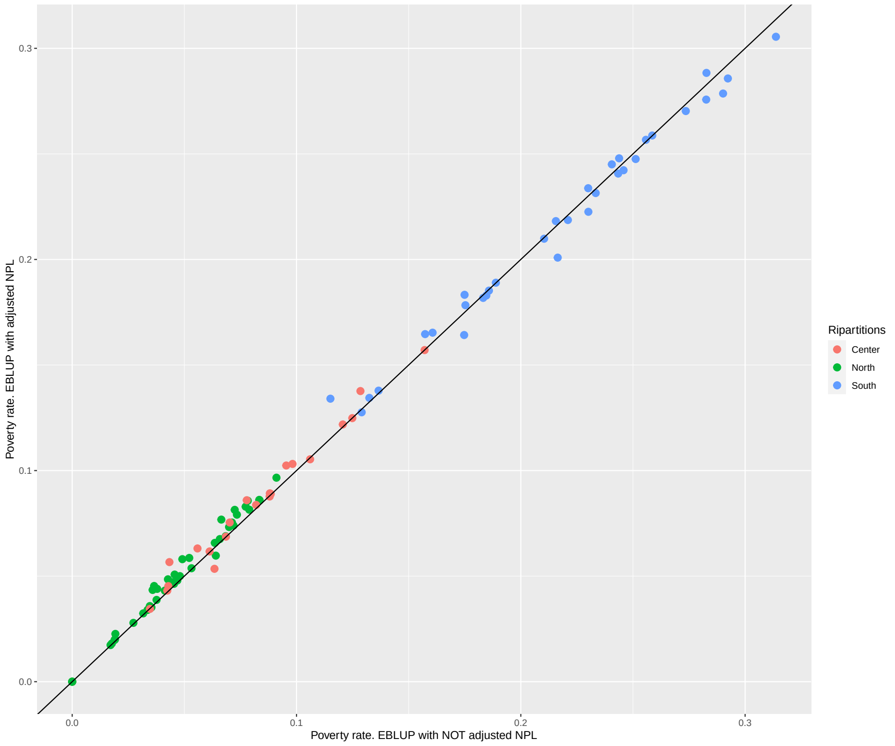


Figure 2: Poverty rate at provincial level in Italy: provincial EBLUPs estimates using the SPI($Q_{0.2}$) adjusted vs not adjusted national poverty line.

References

Laureti, T. and Rao, D. (2018). Measuring spatial price level differences within a country: Current status and future developments. *Estudios de economia aplicada*, 36(1):119–148.

Renwick, T., Aten, B., Figueroa, E., and Martin, T. (2014). Supplemental poverty measure: A comparison of geographic adjustments with regional price parities vs. median rents from the american community survey. Technical report, Bureau of Economic Analysis.

Suits, D. (1984). Dummy variables: Mechanics v. interpretation. *Review of Economics and Statistics*, 66:177–180.

World Bank Group (2015). *Operational Guidelines and Procedures for Measuring the Real Size of the World Economy.* 2011 International Comparison Program, Washington, DC.

Informative Selection and Spatial Processes

Daniel B. Bonnéry [a]*, Francesco Pantalone [b]** and M. Giovanna Ranalli [c]***

*Epidemiology and Modelling Group, Department of Plant Sciences, University of Cambridge, UK.
**Department of Economics, University of Perugia, Italy.
***Department of Political Science, University of Perugia, Italy.

Abstract

This paper extends the concepts of informative selection, population distribution and sample distribution to a spatial process context. These notions were first defined in a context where the output of the random process of interest consists of independent and identically distributed realisations for each individual of a population. It has been showed that informative selection induces stochastic dependence among realisations on selected units. In the context of spatial processes, the "population" is a continuous space and realisations for two different elements of the population are not independent. We show how informative selection may induce a different dependence among selected units from a spatial process and how the sample distribution differs from the population distribution. We provide the correct likelihood and predictive distribution that account for the informative selection on a particular case in this paper, to illustrate the general framework developed by the authors.

Keywords— Nuisance parameters, endogeneous selection, variogram.

1 Introduction

Informative selection occurs when the sampling (selection) process is dependent on the process of interest being measured. This dependence must be accounted for when inferring on the distribution of the study process. When the study process can be described with a simple linear model, where all units of a population behave independently conditionally on the covariates, Skinner et al. (1989) showed that ignoring the selection process can lead to bias and erroneous inference. Krieger and Pfeffermann (1992) distinguish the population distribution from the sample distribution: the observations on

[a] dbb31@cam.ac.uk
[b] francesco.pantalone@studenti.unipg.it
[c] giovanna.ranalli@unipg.it

the sample follow a distribution that is different from the distribution of the units in the population. Bonnéry et al. (2012) also show that, in addition, informative selection can introduce a dependence between the sampled units. When the study process is a spatial process on a field, the population can be assimilated to the field and there is a dependence structure on the population units. In certain cases, this dependence can be summarised by the variogram, see Cressie (2015). In this study, we show that informative selection results in making the dependence structure of the sampled units different from the dependence structure of the "population" units. We apply some of the general results of Bonnery, Pantalone, Ranalli (2021) to a particular example. In fact, we will focus on a particular distribution for the variable of interest and a specific sampling design defined by a Binomial point process. Section 2 introduces the statistical framework. In Section 3 we compare the sample and the population likelihoods when the target is a function of a spatial process sampled informatively. Section 4 focuses on the semivariogram as a relevant example of target of interest: we show the bias that affects the naive sample semivariogram in presence of informative sampling and develop a prediction approach that accounts for it.

2 Statistical Framework

We consider the space $\mathrm{U} = [0,1]^2$ to be our population of interest and a 0-mean isotropic second order stationary random Gaussian process Y with Gaussian covariogram defined on U with value in $\mathscr{Y} = \mathbb{R}$, e.g. $Y : \Omega \to (\mathrm{U} \to (\mathbb{R}, \eta))$, where η is the Lebesgue measure on $\mathscr{Y}$, and ν is a probability measure on U. Capital letters are used for random variables, and corresponding bold lowercase letters for realisations. The process Y is characterised by its finite dimensional densities: for a sample of locations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of size n, $Y[\mathbf{x}]$ denotes the vector $(Y[\mathbf{x}_1], \ldots, Y[\mathbf{x}_n])$. The density of $Y[\mathbf{x}]$ is $\mathrm{f}_{Y[\mathbf{x}]}(\mathbf{y}) = \left(2\pi^{n/2}|\Sigma_Y|^{\frac{1}{2}}\right)^{-1} \exp\left(-\frac{1}{2}\mathbf{y}\Sigma_Y^{-1}\mathbf{y}^{\mathrm{T}}\right)$, where Σ_Y is the matrix with ij-th element given by $\sigma_Y^2 \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\delta_Y^2)$, with σ_Y and δ_Y positive constants. The function $C(h;\sigma_Y,\delta_Y) = \sigma_Y^2 \exp(-\|h^2\|/\delta_Y^2)$, is the covariogram of Y. Figure 2 contains heatmaps of realisations of such processes for different values of the scale parameter δ_Y, respectively 0.01, 0.1 and 1.

The process Y is not observed entirely, but only on a sample. The sampling design is a function of a design variable Z, a random process defined on U with values in $\mathbb{R}$, characterised by $Z = \exp(\alpha_0 + \alpha_1 Y + \alpha_2 \varepsilon)$, where α_0, α_1, α_2 are real positive numbers, and $\varepsilon : \Omega \to (\mathrm{U} \to \mathbb{R})$ is an isotropic Gaussian process with mean $\mu_\varepsilon[\mathbf{x}] = 0$ and Gaussian covariogram $C(h) = \exp\left(-\|h\|^2/\delta_\varepsilon^2)\right)$ independent on Y. The sample S is a random point process on U, such that the distribution of S conditionally on $Z = \mathbf{z}$ is a binomial point process of size n with intensity proportional to $\mathbf{z}$ (as represented in Figure 2). The process S does not depend on Y conditionally on Z and has density with respect to $\nu^{\otimes\{1,\ldots,n\}}$: for $\mathbf{x} \in \mathrm{U}^{\{1,\ldots,n\}}$, $\mathrm{f}_{S|Z=\mathbf{z}}(\mathbf{x}) = \left(\int_U \mathbf{z}(\mathbf{x}')\mathrm{d}\nu(\mathbf{x}')\right)^{-n} \left(\prod_{\ell=1}^n \mathbf{z}(\mathbf{x}_\ell)\right)$.

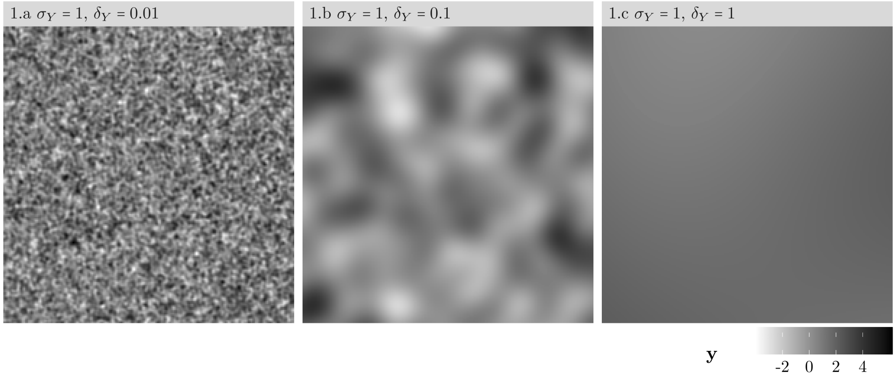


Figure 1: Heat maps of realisations of Y for different values of δ_Y.

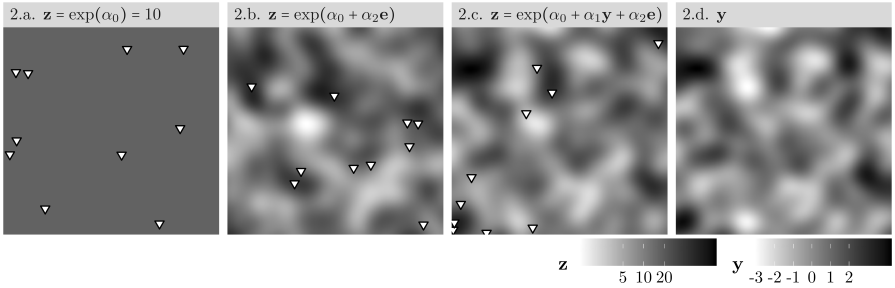


Figure 2: Heat maps of the design variable $\mathbf{z}$ and plot of realisations of S. Sub-figures correspond to the heatmaps of three different design variables $\mathbf{z}$ (Sub-figures 2.a, 2.b, 2.c) and of $\mathbf{y}$ (Sub-figure 2.d) with samples (triangle dots) drawn accordingly to $\mathbf{z}$. The values of $(\alpha_0, \alpha_1, \alpha_2)$ for each sub-figure are: 2.a: $(\log(10), 0, 0)$, 2.b:$(\log(10) - 0.125, 0, 0.5)$, 2.c: $(\log(10) - 0.125, 0.4, 0.3)$

3 Sample vs Population distribution of Y

The data consist of the realisations of the random variables S and $Y[S]$. In this section, we derive the distribution of the observed values of Y on the sample, i.e. the distribution of $Y[S]$, from the distribution of the design variable Z conditionally to the signal Y and the function that links the design to the design variable, or equivalently the distribution of the sample S conditionally to the design variable Z. We resort to the Bayes formula to express the density of $(Y[S], S)$ in $(\mathbf{y}, \mathbf{x})$, as the product of the density of the sample S in $\mathbf{x}$ conditionally on $Y[\mathbf{x}] = \mathbf{y}$ by the density of the signal $Y[\mathbf{x}]$ in $\mathbf{y}$. For $1 \leq n' \leq n$, let $S_{\{1,\ldots,n'\}}$ be the vector $(S[1], \ldots, S[n'])$. For $\mathbf{x} \in \mathrm{U}^{\{1,\ldots,n'\}}$, $\mathbf{y} \in \mathscr{Y}^{\{1,\ldots,n'\}}$,

$$\mathrm{f}_{Y[S_{\{1,\ldots,n'\}}] | S_{\{1,\ldots,n'\}}}(\mathbf{y} \mid \mathbf{x}) = \rho_{\{1,\ldots,n'\}}(\mathbf{x} \mid \mathbf{y}) \times \mathrm{f}_{Y[S_{\{1,\ldots,n'\}}]}(\mathbf{y}) \neq \mathrm{f}_{Y[\mathbf{x}]}(\mathbf{y}), \qquad (1)$$

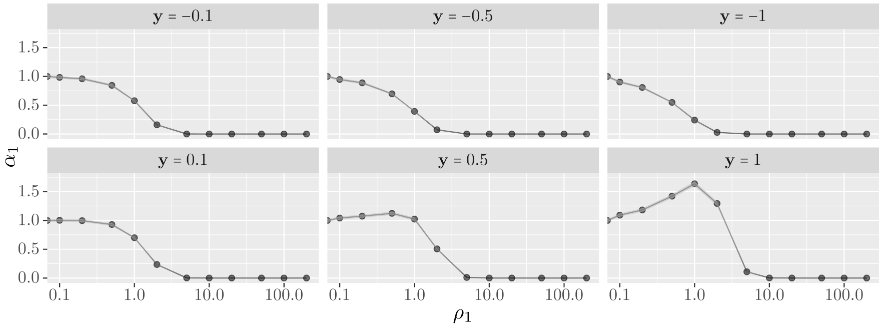


Figure 3: Representation of $\tilde{\rho}_{\{1\}}(\mathbf{x} \mid \mathbf{y})$ when $\alpha_2 = 1$ and α_1 varies.

where

$$\rho_{\{1,\ldots,n'\}}(\mathbf{x} \mid \mathbf{y}) \\ = \frac{\int\prod_{\ell=1}^{n'}(\int \exp\left(\alpha_1 (Y[\mathbf{x}'] - \mathbf{y}_\ell) + \alpha_2 (\varepsilon[\mathbf{x}'] - \varepsilon[\mathbf{x}_\ell])\right) \mathrm{d}\nu(\mathbf{x}')\mathrm{d}\nu(\mathbf{x}')\big)^{-1} \mathrm{d}P^{Y,\varepsilon|Y[\mathbf{x}]=\mathbf{y}}}{\int \left(\int \exp\left(\alpha_1 (Y[\mathbf{x}'] - Y[\mathbf{x}_\ell]) + \alpha_2 (\varepsilon[\mathbf{x}'] - \varepsilon[\mathbf{x}_\ell])\right) \mathrm{d}\nu(\mathbf{x}')\mathrm{d}\nu(\mathbf{x}')\right)^{-n'} \mathrm{d}P^{Y,\varepsilon}}.$$

Figure 3 represents the function $\rho_{\{1\}}(\mathbf{x} \mid \mathbf{y})$ for $\mathbf{x} = (0.5, 0.5)$ and when $\mathbf{y}$ and the nuisance parameters α_1 vary and the other parameters are fixed. Numerical methods are used to compute $\rho_{\{1,\ldots,n\}}$, which allows to compute the likelihood of the different parameters based on Equation (1).

4 Estimation of semivariogram parameters and prediction

Figure 4 shows that in the case of informative selection ($\alpha_1 \neq 0$), the naive estimator of the semivariogram is biased. The couple of spatial process Y and ε was simulated one time independently, and for three different values of the nuisance parameter vector $\alpha = (\alpha_0, \alpha_1, \alpha_2)$, 1000 samples of sample size $n = 100$ were drawn. The solid line represents the semivariogram obtained from all the points of U, whereas the dashed line is the pointwise mean of all the semivariogram estimates. When α_1 is large, it has the effect of skewing the sample distribution of Y, which in turns has the effect of producing naive estimates that overestimate the variance of Y and the sill, i.e. limit when $h \to \infty$.

This is an illustration of the fact that the expected value of the naive nonparametric estimate of the semivariogram is given by approximately

$$\frac{1}{2}\int_{\mathrm{U}^{\{1,2\}}} \left[\int_{\mathscr{Y}^{\{1,2\}}} (\mathbf{y}_2 - \mathbf{y}_1)^2 \, \rho_{\{1,2\}}(\mathbf{x} \mid \mathbf{y}) \, \mathrm{f}_{Y[\mathbf{x}]}(\mathbf{y}) \, \mathrm{d}\eta^{\otimes 2}(\mathbf{y})\right] \mathrm{d}(\nu^{\otimes\{1,2\}})^{X|X_2 - X_1 = h}(\mathbf{x})$$

which differs by the term $\rho_{\{1,2\}}$ from the value of the semivariogram of Y:

$$\frac{1}{2}\int_{\mathrm{U}^{\{1,2\}}} \left[\int_{\mathscr{Y}^{\{1,2\}}} (\mathbf{y}_2 - \mathbf{y}_1)^2 \, \mathrm{f}_{Y[\mathbf{x}]}(\mathbf{y}) \mathrm{d}\eta^{\otimes 2}(\mathbf{y})\right] \mathrm{d}(\nu^{\otimes\{1,2\}})^{X|X_2 - X_1 = h}(\mathbf{x}).$$

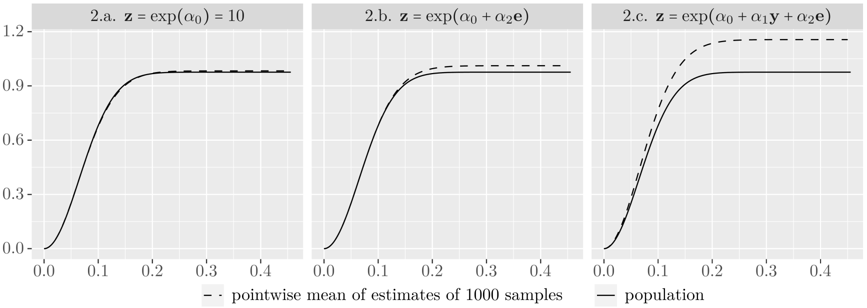


Figure 4: Naive estimates of the semivariogram.

To account for the informative selection process, we propose to base our estimates on the true conditional likelihood:

$$L(\alpha, \sigma_Y, \delta_Y, \delta_\varepsilon; (Y[\mathbf{x}] = \mathbf{y} \mid S = \mathbf{x})) = \rho_{\{1,\ldots,n\};\alpha,\sigma_Y,\delta_Y,\delta_\varepsilon}(\mathbf{x} \mid \mathbf{y}) \times \mathrm{f}_{Y[S];\sigma_Y,\delta_Y}(\mathbf{y}).$$

In addition, we can also consider prediction accounting for informative selection. In particular, the following equation characterises the predictive distribution of $Y[\mathbf{x}_0]$ when $S, Y[S] = \mathbf{x}, \mathbf{y}$ is observed that takes into account the informative selection process: the probability density function of $Y[\mathbf{x}_0]$ conditionally on $(S, Y[S]) = (\mathbf{x}, \mathbf{y})$ is given by:

$$\mathrm{f}_{Y[\mathbf{x}_0]|S,Y[S]}(\mathbf{y}_0 \mid \mathbf{x}, \mathbf{y}) = \frac{\mathrm{f}_{S|Y[\mathbf{x}],Y[\mathbf{x}_0]}(\mathbf{x} \mid \mathbf{y}, \mathbf{y}_0)}{\mathrm{f}_{S|Y[\mathbf{x}]}(\mathbf{x} \mid \mathbf{y})} \times \mathrm{f}_{Y[\mathbf{x}_0]|Y[\mathbf{x}]}(\mathbf{y}_0 \mid \mathbf{y}), \tag{2}$$

with

$$\frac{\mathrm{f}_{S|Y[\mathbf{x}],Y[\mathbf{x}_0]}(\mathbf{x} \mid \mathbf{y}, \mathbf{y}_0)}{\mathrm{f}_{S|Y[\mathbf{x}]}(\mathbf{x} \mid \mathbf{y})} = \frac{\int \frac{\exp\left(-\alpha_1 (\sum_{\ell=1}^{n'} \mathbf{y}(\ell) - \alpha_2 (\sum_{\ell=1}^{n'} \varepsilon[\mathbf{x}_\ell])\right)}{(\int \exp(-\alpha_1 Y[\mathbf{x}_0] - \alpha_2 \varepsilon[\mathbf{x}_0]) \mathrm{d}\nu(\mathbf{x}_0))^{n'}} \mathrm{d}P^{Y,\varepsilon|Y[\mathbf{x}]=\mathbf{y},Y[\mathbf{x}_0]=\mathbf{y}_0}}{\int \frac{\exp\left(-\alpha_1 (\sum_{\ell=1}^{n'} Y[\mathbf{x}_\ell] - \alpha_2 (\sum_{\ell=1}^{n'} \varepsilon[\mathbf{x}_\ell])\right)}{(\int \exp(-\alpha_1 Y[\mathbf{x}_0] - \alpha_2 \varepsilon[\mathbf{x}_0]) \mathrm{d}\nu(\mathbf{x}_0))^{n'}} \mathrm{d}P^{Y,\varepsilon|Y[\mathbf{x}]=\mathbf{y}}}.$$

Naive prediction would use $\mathrm{f}_{Y[\mathbf{x}_0]|Y[\mathbf{x}]}(\mathbf{y}_0 \mid \mathbf{y})$, which leads to biased predictions as well as biased estimation of the confidence interval of predictions. Via numerical methods, when model parameters are known or plugged in, we integrate expression (2) to compute point predictors $E[Y[\mathbf{x}_0] \mid S = \mathbf{x}, Y[S] = \mathbf{y}]$ as well as block predictors $E[\int_{A_0} Y[\mathbf{x}'] \mathrm{d}\nu(\mathbf{x}') \mid S = \mathbf{x}, Y[S] = \mathbf{y}]$, their variance and their confidence intervals.

5 Conclusion

We have extended the notions of informative selection, population and sample distributions defined by Krieger and Pfeffermann (1992) to a situation where a spatial process is assumed to have generated the whole population. In particular, we have focused on a Gaussian random field and a binomial point process that represents the selection mechanism.

We have highlighted the role of the density ratio function ρ and expressed correct likelihood and predictive distribution to account for the informative selection process. We have focused on estimation of the semivariogram and highlighted the bias coming from naive estimation in presence of informative selection. We use numerical methods to estimate the function ρ, but they can be computationally intensive. Alternative methods that use approximations are currently being investigated.

Acknowledgements Francesco Pantalone's work was partially funded by the *International Graduate Research Fellowships* at Joint Program in Survey Methodology, University of Maryland, College Park, USA.

References

Bonnéry, D., Breidt, F. J., Coquet, F., et al. (2012). Uniform convergence of the empirical cumulative distribution function under informative selection from a finite population. *Bernoulli*, 18(4):1361–1385.

Bonnery, D., Pantalone, F., and Ranalli, M. G. (2021). The effect of informative selection on the estimation of parameters related to spatial processes. *arXiv preprint arXiv:2103.10540*.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Krieger, A. M. and Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology*, 18(2):225–239.

Skinner, C. J., Holt, D., and Smith, T. F. (1989). *Analysis of complex surveys*. John Wiley & Sons.

Estimation of life expectancy in small areas using big data from the municipal registry

Stefano Cervellera [a*], Carlo Cusatelli [b**] and Massimiliano Giacalone [c***]

[*]Municipality of Taranto
[**]Ionian Department, University of Bari 'Aldo Moro'
[***]Department of Economics and Statistics, University of Naples 'Federico II'

Abstract

Given the importance of knowing life expectancy, with less aggregation than provincial level, the paper aims to estimate it for municipal and sub-municipal areas, by means of big data from Registry of Municipal Census Office. The UN approved the 2030 Agenda, with 169 Sustainable Development Goals in 17 domains, placing in SDG 3 (Health and well-being) as many as 6 goals for mortality reduction out of 9 total and for improvement of life expectancy at birth and healthy life. We highlight the validity of a tool, such as the real-time mortality observatory, which we have been designing for some years as a support to both medicine and criminal justice, as an alternative to the classic epidemiological analysis and appraisals based on outdated information. For Taranto city, where the environmental pressure is very strong, different statistical sources can be used to shorten these waiting times compared to ecological and epidemiological studies based on official data which have a longer circuit.

Keywords— Life expectancy, Mortality, Istat, Registry-office, Taranto)

1 Introduction

Mortality tables are one of the most complete tools in determination of biometric functions: deaths, probability of death and survival, years lived and life expectancy. However, their elaboration requires specific data on the structure of the population and on the characteristics of deaths. In Italy, Istat has always placed considerable

[a]s.cervellera@comune.taranto.it
[b]carlo.cusatelli@uniba.it
[c]massimiliano.giacalone@unina.it

interest in their construction since 1972, with regional aggregates, and since 1995 at the provincial level, in line with the European harmonization on the nomenclature levels of the NUTS 3 statistical territorial units (from French "Nomenclature des unités territoriales statistiques", it includes regions with populations between 150,000 and 800,000 people, for example the oblasts in Bulgaria and the provinces in Italy).
From the official statistics, life expectancy by age can be obtained from the ISTAT mortality tables (Demo.istat datawarehouse) from 1974 to 2020, for NUTS 3 provincial aggregates (107 today in Italy), with an average population of about 550,000 inhabitants. But the relevance of mortality and life expectancy data for smaller municipal and sub-municipal areas is increasingly greater in terms of awareness and active citizenship, as well as for governments and organizations (at any level, including international), in intervention policies on issues concerning the health of citizens (Gianicolo et al., 2016; Vigotti et al., 2014).

2 Life expectancy in the Taranto province

In 2006 Taranto was among the top ten provinces with the best life expectancy, equal to almost 82 years (1), but since then it has seen a sharp decline, recovering this level only five or six years later: in that period we were among the worst in Italy. In 2009 we had lost 70 positions, a situation never seen before: only in the war phase this can happen, and we wondered why and how it happened in such a short time, so much so that at the end of this phase the judicial processes began on the many environmental problems of Taranto. Analysing the most easily accessible source of data, that of the National Institute of Statistics, we realize that years of delay are also accumulated for the validation of causes of death, thus making its databases provisional for a long time. Even at the local level, the latest cancer register relating to Taranto is from December 2017, an excellent job, which however presents data updated to 2012, therefore, with a large time gap that requires checking if there are alternatives. In particular for the city of Taranto, where the environmental pressure is actually very strong, different statistical sources can be used to shorten these waiting times compared to ecological and epidemiological studies based on official data, which have a

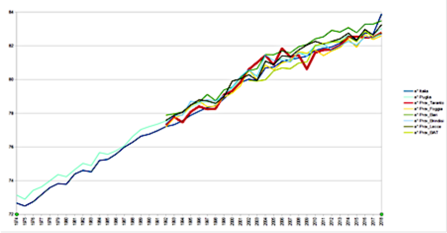


Figure 1: Life expectancy at birth, e^o (Source: Istat)

longer circuit (Berti et al., 2009): having available a direct source, such as the municipal registry, it was possible for us to make even sub-municipal analysis, whereas Istat provides mostly provincial data.

Working on that aggregate basis is a limit, and we realize it especially in this phase of the Covid-19 pandemic, in which there is the frantic search for local and up-to-date data, where the measurement system must be correct, precise and updated as much as possible, for the control system it has to enslave. If you intend to use demographic indices, such as those relating to mortality, to develop health policies, you need to use the data immediately: a slow measurement system does not work, it must be speeded up as much as possible.

3 Life expectancy in Taranto city and sub-municipal areas

We therefore consulted the aforementioned data sources and we were able to understand that the trend of Istat presented, above all in the past, large differences compared to the municipal registry of Taranto (2): especially in the first decade of the new millennium, therefore between the fourteenth and fifteenth Census, there were quite misaligned numbers between Istat (whose data trend is rather irregular) and registry office (with a fairly regular evolution). In Tab. 1 we can therefore see the difference in life expectancy between males and females, in the individual districts of Taranto. Also Istat is now opening up to the registry data, a good result for the whole community, so in January 2019 we had the publication of life expectancy for all Italian cities. Fig. 3 graphically represents those data: life expectancy grows a little in all the districts, among which there is however a lot of inequality. For example, in the Tre Carrare-Solito area, the life expectancy of men and women in 2018 was almost the same, and this makes the trend of the latter even negative; but even in some circumscriptions there is a paradoxically lower female life expectancy than male even though it concerns different but still close neighborhoods.

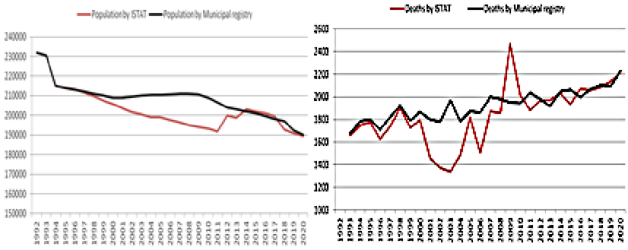


Figure 2: (Mis)alignments Registry-Office vs Istat

From the indicators of life expectancy at birth, a situation of strong inequality emerges, which obviously affects the socio-economic situation: the Borgo, Paolo

Table 1: Life expectancy at birth, by year, gender and district

Gender	District	Year										
		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Females	Paolo VI	81.43	82.60	84.27	83.94	82.84	80.86	81.86	83.40	82.56	84.13	83.32
	Tamburi	83.52	82.42	84.15	83.74	83.63	84.08	85.19	81.54	84.57	84.45	83.71
	Borgo	83.51	83.45	82.86	84.19	84.00	84.85	83.13	83.65	84.23	84.15	84.43
	Tre Carrare - Solito	84.72	84.71	85.31	85.36	85.59	86.50	84.20	84.70	83.69	85.63	85.34
	Montegranaro - Salinella	83.99	85.98	85.18	86.30	85.00	86.03	85.46	86.45	84.58	84.93	85.41
	Talsano Lama S. Vito	84.96	83.83	85.45	85.69	83.89	85.32	86.06	84.89	87.00	85.40	85.15
Males	Paolo VI	77.71	79.24	77.79	79.61	78.66	79.88	78.07	79.36	78.63	74.95	79.71
	Tamburi	75.35	76.85	78.08	78.84	77.78	77.31	78.46	79.80	77.37	77.42	77.48
	Borgo	77.79	76.98	77.02	79.00	78.38	74.87	77.42	78.74	79.39	80.09	78.82
	Tre Carrare - Solito	81.46	80.01	79.89	80.82	80.31	81.53	80.19	82.36	83.61	82.18	82.69
	Montegranaro - Salinella	81.08	79.91	80.23	80.67	80.82	81.88	81.62	81.98	81.69	82.86	81.60
	Talsano Lama S. Vito	81.25	79.65	80.72	81.12	81.60	81.36	81.31	80.89	81.02	83.56	82.79

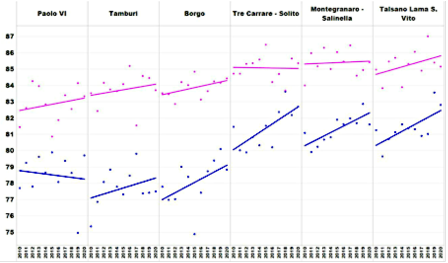


Figure 3: Life expectancy at birth, by year, gender and district

VI and Tamburi districts are united by the levels of life expectancy very detached from the others, as if they were different cities (Fig. 4), which also results from life expectancy at all ages (Fig. 5).
In fact, without considering causes of death, we also calculated Standardized Mortality Ratios (SMR) between observed and expected deaths estimated by taking Apulian mortality rates as a reference: the map in Fig. 6 can therefore be interpreted considering the area of residence as an exposure factor (Graziano et al., 2009; Mangia et al., 2013; Marinaccio et al., 2011; Mataloni et al., 2012). For example, we record a statistically significant excess of 21% for Tamburi in the period 2010-2018. But we also have positive data that show how resilient the city is: the over one-hundred-year olds are also distributed in the areas of greatest environmental impact, with people aged 109, 110 and even 112 years old.

4 Final remarks

This kind of analysis would also make it possible to verify generational inequalities in small areas and, due to the ever faster updating of the municipal registries, the data could be processed in near-real-time. In a future work we will determine the ID-Sentieri socio-deprivation indicators for municipal and sub-municipal areas,

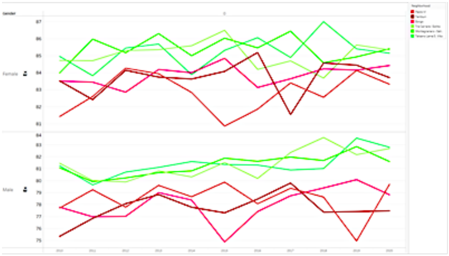


Figure 4: Life expectancy at birth, by year, gender and district

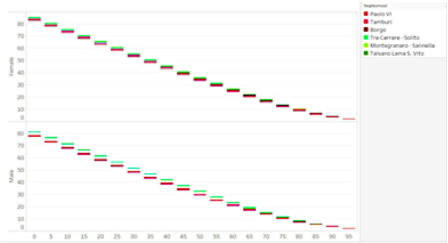


Figure 5: Life expectancy by age, gender and district (average 2010-2020)

which in the literature is related to indicators such as the standardized mortality rate.

References

Berti, G., Galassi, C., Faustini, A., and Forastiere, F. (2009). Inquinamento atmosferico e salute, sorveglianza epidemiologica e interventi di prevenzione. *Epidemiol Prevenzione*, Suppl 1:1–144.

Gianicolo, E., Mangia, C., and Crevino, M. (2016). Investigating mortality heterogeneity among neighborhoods of a highly industrialized Italian city: a meta-regression approach. *Int J Public Health* , 61:777–785.

Graziano, G., Bilancia, M., Bisceglia, L., de Nichilo, G., Pollice, A., and Assennato, G. (2009). Statistical analysis of the incidence of some cancers in the province of Taranto 1999-2001. *Epidemiol Prevenzione*, 33:37–44.

Mangia, C., Gianicolo, E.A., Bruni, A., Vigotti, M.A., and Cervino, M. (2013). Spatial variability of air pollutants in the city of Taranto, Italy and its potential impact on exposure assessment. *Environ Monit Assess.*, 185:1719–1735.

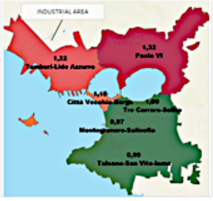


Figure 6: SMR (average 2011-2019) in significant excess (red areas) and defect (green areas) compared to Apulian mortality, with a 90% confidence level

Marinaccio, A., Belli, S., and Binazzi, A. (2011). Residential proximity to industrial sites in the area of Taranto (Southern Italy). A case-control cancer incidence study. In *Ann Ist Super Sanitá*, 47:191–199.

Mataloni, F., Stafoggia, M., Alessandrini, E., Triassi, M., Biggeri, A., and Forastiere, F. (2012). A cohort study on mortality and morbidity in the area of Taranto, Southern Italy. *Epidemiol Prevenzione*, 36(5):237–252.

Vigotti, M.A., Mataloni, F., Bruni, A., Minniti, C., and Gianicolo, E. A. (2014). Mortality analysis by neighborhood in a city with high levels of industrial air pollution. *Int J Public Health* , 59(4):645–653.

Design-based composite estimation of small proportions in small domains

Andrius Čiginas [a]*

*Faculty of Mathematics and Informatics, Vilnius University

Abstract

Traditional direct estimation methods are not efficient for domains of a survey population with small sample sizes. To estimate the domain proportions, we combine the direct estimators and the regression-synthetic estimators supported by domain-level auxiliary information. For the case of small true proportions, we introduce the design-based linear combination that is a robust alternative to the empirical best linear unbiased predictor (EBLUP) based on the Fay–Herriot model. We imitate the Lithuanian Labor Force Survey, where we estimate the proportions of the unemployed in municipalities. We show that the considered design-based compositions and estimators of their mean square errors are competitive for EBLUP and its accuracy estimation.

Keywords— area-level model, composite estimator, Labor Force Survey

1 Introduction

Design-based and model-assisted direct estimators of parameters rely only on the sample of the estimation domain (area). Therefore, after the sample is selected, their application for some unplanned domains leads to high variances of the estimators because of too small sample sizes. In the small area estimation theory (Rao and Molina, 2015), indirect estimators borrow sample information from neighbor domains through auxiliary information and linking models. These model-based estimators usually have lower variances than the direct estimators, but their biases can be significant.

To estimate proportions in the domains, one can consider explicit linking models supported by auxiliary data aggregated to the area level. A popular model is the Fay–Herriot (FH) model, which is a separate case of linear mixed models, and the empirical best linear unbiased predictors (EBLUPs) of the domain means or proportions are derived from it (Fay and Herriot, 1979).

[a] andrius.ciginas@mif.vu.lt

That small area predictor is expressed as the linear combination of a regression-synthetic estimator and the direct estimator. While the former part accounts for a variation reflected in the auxiliary data, the direct component exploits the unbiasedness property. Compositions of the synthetic and the direct estimators constitute an important class of indirect estimators. Before the mixed models, traditional design-based composite estimators were often used (Rao and Molina, 2015, Chapter 3). However, now it is accepted that the models including random area-specific effects are more useful. For example, they are more convenient to handle complex data structures than the traditional estimators with only randomness induced by the sampling design. Another notable drawback of the latter estimators is the difficulty to estimate their precision.

We construct the design-based composite estimator, which is in some sense similar to EBLUP. According to the construction, it is a robust estimator suitable for small or large domain proportions. The mean square error (MSE) of this composition is estimated as suggested in Čiginas (2021). We compare the estimators and their MSE estimators in the simulation study using the Lithuanian Labor Force Survey (LFS) data, where fractions of the unemployed are the proportions of interest estimated in municipalities.

2 Direct and synthetic estimation

The set $\mathcal{U} = \{1, \ldots, N\}$ consists of the labels of elements of the survey population. Let y be a binary study variable with the fixed values $y_1, \ldots, y_N$ assigned to the corresponding elements. The sample $s \subset \mathcal{U}$ of size $n < N$ is drawn by the sampling design $p(\cdot)$, and $\pi_k = \mathrm{P}_\mathrm{p}\{k \in s\} > 0$, $k \in \mathcal{U}$, are inclusion into the sample probabilities. Here the symbol P_p, and hereafter E_p, var_p, and MSE_p denote probability, expectation, variance, and MSE according to $p(\cdot)$, respectively. The characteristic $\mathrm{var}_\mathrm{p}(\cdot)$ is called the sampling variance or design variance. Let $\mathcal{U} = \mathcal{U}_1 \cup \cdots \cup \mathcal{U}_M$ be the partition of the population into the non-overlapping domains, where the domain $\mathcal{U}_i$ contains N_i elements. Then the domain sample $s_i = s \cap \mathcal{U}_i$ is of size $n_i \leq N_i$. We aim to estimate the proportions $\theta_i = \sum_{k \in \mathcal{U}_i} y_k / N_i$, $i = 1, \ldots, M$, where the numbers N_i are assumed to be known. If the design $p(\cdot)$ does not ensure the fixed sizes n_i, then they can be too small to get suficiently accurate direct estimates $\hat{\theta}_i^\mathrm{d}$ of θ_i.

Assume that, for each domain $\mathcal{U}_i$, the auxiliary information is available as the vector of characteristics $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iP})'$. This assumption narrows a choice of direct estimators to the design unbiased Horvitz–Thompson estimators of θ_i or the weighted sample proportions

$$\hat{\theta}_i^\mathrm{d} = \frac{1}{\widehat{N}_i} \sum_{k \in s_i} \frac{y_k}{\pi_k}, \quad \text{where} \quad \widehat{N}_i = \sum_{k \in s_i} \frac{1}{\pi_k}, \qquad i = 1, \ldots, M, \tag{1}$$

that are approximately unbiased. Let $\hat{\psi}_i$ be estimators of the sampling variances $\psi_i = \mathrm{var}_\mathrm{p}(\hat{\theta}_i^\mathrm{d})$. The direct estimators $\hat{\psi}_i = \hat{\psi}_i^\mathrm{d}$ (Särndal et al., 1992, p. 185) are approximately design unbiased but they have large variances themselves for small sample sizes. Therefore, the variances $\hat{\psi}_i^{\mathrm{d}}$ are smoothed, and new more stable estimators $\hat{\psi}_i^{\mathrm{s}}$ are often further used. For the proportions, that $\hat{\psi}_i^{\mathrm{s}}$ can be obtained by assuming that $\psi_i \approx K N_i^{\gamma}$ and estimating the parameters $K > 0$ and $\gamma \in \mathbb{R}$ through a log-log regression model.

The regression-synthetic estimators

$$\hat{\theta}_i^{\mathrm{S}} = \hat{\theta}_i^{\mathrm{S}}(\hat{\psi}_i) = \mathbf{z}_i' \hat{\boldsymbol{\beta}} \quad \text{with} \quad \hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{M} \frac{\mathbf{z}_i \mathbf{z}_i'}{\hat{\psi}_i} \right)^{-1} \sum_{i=1}^{M} \frac{\mathbf{z}_i \hat{\theta}_i^{\mathrm{d}}}{\hat{\psi}_i}, \qquad i = 1, \ldots, M, \quad (2)$$

of θ_i are derived from the basic area-level model for EBLUP ignoring random area effects (Rao and Molina, 2015, Section 4.2). Here one can take $\hat{\psi}_i = \hat{\psi}_i^{\mathrm{s}}$ instead of $\hat{\psi}_i^{\mathrm{d}}$. If the underlying linking model is strong, the sampling variances of (2) are small, compared to that of $\hat{\theta}_i^{\mathrm{d}}$. However, the design biases of (2) can be relatively large because a specificy of the domains is not taken into account.

3 Design-based composite estimation

3.1 Evaluation of optimal compositions and their accuracy estimation

To find a trade-off between larger variances of $\hat{\theta}_i^{\mathrm{d}}$ and biases of $\hat{\theta}_i^{\mathrm{S}}$, we consider their linear combinations

$$\tilde{\theta}_i^{\mathrm{C}} = \tilde{\theta}_i^{\mathrm{C}}(\lambda_i) = \lambda_i \hat{\theta}_i^{\mathrm{d}} + (1 - \lambda_i) \hat{\theta}_i^{\mathrm{S}}, \qquad i = 1, \ldots, M, \quad (3)$$

with weights $0 \le \lambda_i \le 1$. Minimizing the function $\mathrm{MSE}_{\mathrm{p}}(\tilde{\theta}_i^{\mathrm{C}}(\lambda_i))$ with respect to λ_i, the optimal weight λ_i^* for the domain $\mathcal{U}_i$ is obtained and then approximated using $\lambda_i^* \approx \mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})/(\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{d}}) + \mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}}))$, see Rao and Molina (2015). However, it is difficult to evaluate the quantities $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})$. A common approach to this is to use the representation

$$\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}}) = \mathrm{E}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}})^2 - \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}}) + \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})$$

with $\hat{\theta}_i^{\mathrm{d}}$ assumed unbiased (Rao and Molina, 2015, Section 3.2.5), and then to build an approximately design unbiased estimator

$$\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{S}}) = (\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}})^2 - \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}}) + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}}), \quad (4)$$

where $\hat{\sigma}^2(\cdot)$ is an estimator of the variance $\mathrm{var}_{\mathrm{p}}(\cdot)$. Unfortunately, estimator (4) can be very unstable and take negative values for individual small domains. Therefore, the straightforward estimation of the optimal weights λ_i^* is avoided.

To alternatively evaluate the optimal coefficients for compositions (3), one can set a common weight for all domains and then minimize a total MSE with respect to that weight (Rao and Molina, 2015, Section 3.4.1). A similar approach is to apply James–Stein method (Rao and Molina, 2015, Section 3.4). Another idea is sample-size-dependent estimation (Rao and Molina, 2015, Section 3.3.2).

Estimation of MSEs of the design-based composite estimators like these is known as a difficult problem in the literature (Rao and Molina, 2015, Chapter 3). One general way is to treat the composition $\hat{\theta}_i^{\mathrm{C}} = \tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i)$ as a synthetic estimator and use the estimator

$$\mathrm{mse_u}(\hat{\theta}_i^{\mathrm{C}}) = (\hat{\theta}_i^{\mathrm{C}} - \hat{\theta}_i^{\mathrm{d}})^2 - \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}} - \hat{\theta}_i^{\mathrm{d}}) + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}}) \tag{5}$$

of $\mathrm{MSE_p}(\hat{\theta}_i^{\mathrm{C}})$. However, (5) has the same drawbacks as (4). Another general method is to assume that the estimator $\hat{\theta}_i^{\mathrm{C}}$ defined by (3) approximates the optimal combination $\hat{\theta}_i^{\mathrm{opt}} = \tilde{\theta}_i^{\mathrm{C}}(\lambda_i^*)$ quite well and derive the approximation $\mathrm{MSE_p}(\hat{\theta}_i^{\mathrm{C}}) \approx \lambda_i(1-\lambda_i)\psi_i + \mathrm{var_p}(\hat{\theta}_i^{\mathrm{C}})$ with the empirical version (Čiginas, 2021)

$$\mathrm{mse_b}(\hat{\theta}_i^{\mathrm{C}}) = \hat{\lambda}_i(1-\hat{\lambda}_i)\hat{\psi}_i + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}}), \tag{6}$$

where we would set $\hat{\psi}_i = \hat{\psi}_i^{\mathrm{s}}$. Estimator (6) takes only non-negative values.

3.2 Composition based on a ratio of variances

The sampling variance ψ_i is approximately proportional to the product $\theta_i(1-\theta_i)$. That is, one can use the approximation

$$\psi_i \approx D_i\theta_i(1-\theta_i)/n_i, \tag{7}$$

where D_i is the design effect reflecting the sample efficiency of the complex sampling design (Kish, 1995). Then, inserting $\hat{\theta}_i^{\mathrm{d}}$ and an appropriate estimator $\widehat{D}_i$ of D_i into (7), we approximate the direct estimator $\hat{\psi}_i^{\mathrm{d}}$ of ψ_i.

Let us first suppose that the domain proportions θ_i are small, say $\theta_i < 0.1$. Consider two candidate estimators $\hat{\psi}_i^{\mathrm{d}}$ and $\hat{\psi}_i^{\mathrm{s}}$ of ψ_i used in (2). Assume that we got too small estimate $\hat{\theta}_i^{\mathrm{d}}$ of θ_i for the specific sample s. The direct estimate $\hat{\psi}_i^{\mathrm{d}}$ then underestimates ψ_i. Therefore, the inequality $\hat{\psi}_i^{\mathrm{s}} > \hat{\psi}_i^{\mathrm{d}}$ should often hold, that is, the smoothed variance $\hat{\psi}_i^{\mathrm{s}}$ could be a better choice than $\hat{\psi}_i^{\mathrm{d}}$. Now suppose that $\hat{\theta}_i^{\mathrm{d}}$ overestimated the parameter θ_i. Then $\hat{\psi}_i^{\mathrm{d}}$ overestimates ψ_i as well, and the inequality $\hat{\psi}_i^{\mathrm{s}} < \hat{\psi}_i^{\mathrm{d}}$ should hold if $\hat{\theta}_i^{\mathrm{d}}$ is an outlier. That larger estimate $\hat{\psi}_i^{\mathrm{d}}$ can be employed to down-weight the outlying observation $\hat{\theta}_i^{\mathrm{d}}$ used in $\hat{\boldsymbol{\beta}}$ thus robustifying estimators (2). From these considerations, we derive the combined estimators $\hat{\psi}_i^{\mathrm{c}} = \max\{\hat{\psi}_i^{\mathrm{s}}, \hat{\psi}_i^{\mathrm{d}}\}$ of ψ_i that should improve the regression-synthetic estimation. Next, we define the composite estimators

$$\hat{\theta}_i^{\mathrm{C}} = \hat{\lambda}_i\hat{\theta}_i^{\mathrm{d}} + (1-\hat{\lambda}_i)\hat{\theta}_i^{\mathrm{S}}(\hat{\psi}_i^{\mathrm{c}}) \quad \text{with} \quad \hat{\lambda}_i = \frac{\min\{\hat{\psi}_i^{\mathrm{s}}, \hat{\psi}_i^{\mathrm{d}}\}}{\hat{\psi}_i^{\mathrm{c}}}, \qquad i = 1, \dots, M, \tag{8}$$

of the proportions θ_i. If the estimate $\hat{\theta}_i^{\mathrm{d}}$ is an outlier by its small or large value, then relatively more weight is attached to the synthetic part of composition (8).

We use the same arguments to create (8) if the parameters θ_i are not small, but then the inequalities $\max\{\theta_i, \hat{\theta}_i^{\mathrm{d}}\} < 1/2$ or $\min\{\theta_i, \hat{\theta}_i^{\mathrm{d}}\} > 1/2$ must be satisfied. If these inequalities are not valid, the composite estimator is still applicable, but it can be less efficient. The worst scenario here would be a large difference $\theta_i - \hat{\theta}_i^{\mathrm{d}}$ and the relation $\theta_i \approx 1 - \hat{\theta}_i^{\mathrm{d}}$ but those events are rare.

To estimate MSE of design-based composition (8), we apply estimator (6).

4 Simulations using the Labor Force Survey data

Let θ_i be the proportions of the unemployed in the municipalities of Lithuania. We construct the artificial population $\mathcal{U}$ from the single sample data by removing areas with too small fractions of unemployed and then replicating the data of each person the number of times equal to the rounded survey weight. We get $M = 30$ and $N = 1396763$. We draw $R = 10^3$ samples of households of size $n' = 3700$ without replacement with probabilities proportional to household sizes. The selected households are surveyed entirely, and it yields $n \approx 7667$.

We compare the direct estimator $\hat{\theta}_i^{\mathrm{d}}$ from (1), synthetic estimator (2), design-based composition (8), and EBLUP $\hat{\theta}_i^{\mathrm{FH}}$, where $\hat{\psi}_i = \hat{\psi}_i^{\mathrm{s}}$ and the variance of the random area effects is estimated using the method of moments. Moreover, we compare the accuracy of the appropriate MSE estimator for $\hat{\theta}_i^{\mathrm{FH}}$ with that of two MSE estimators (5) and (6) applied to (8). We consider also the optimal combination $\hat{\theta}_i^{\mathrm{opt}}$ and its MSE estimator by (6). The auxiliary proportions in $\mathbf{z}_i = (1, z_{i2}, z_{i3}, z_{i4}, z_{i5}, z_{i6})'$ are: z_{i2} is registered unemployment, z_{i3} means persons who pay the social contribution, z_{i4} is for males, and z_{i5} and z_{i6} are for age intervals 26–40 and 41–55, respectively. We apply the bootstrap to evaluate the variances in (5) and (6). We use the accuracy measures

$$\mathrm{RMSE}(\hat{\mu}_i) = \left(\frac{1}{R}\sum_{r=1}^{R}(\hat{\mu}_i^{(r)} - \mu_i)^2\right)^{1/2} \quad \text{and} \quad \mathrm{AB}(\hat{\mu}_i) = \left|\frac{1}{R}\sum_{r=1}^{R}\hat{\mu}_i^{(r)} - \mu_i\right|,$$

where $\hat{\mu}_i^{(r)}$ is a realization of the specific estimator $\hat{\mu}_i$ of the parameter μ_i, based on the rth sample. We classify the municipalities by the expected domain sample size into three classes of equal size, and calculate the average of RMSEs as well as ABs over domains of each class. We also present the averages over all municipalities as common indicators of accuracy.

The results are in Table 1. We use the superscripts of estimators to discuss the output. Any indirect estimator of the proportions improves the direct one in the sense of RMSE, and theoretical composition opt is the best estimator. Among the indirect estimators, synthetic estimator S has much larger design biases than compositions FH and C. The average of RMSEs over all domains of design-based composite estimator C is smaller than that of FH. MSE estimation (6) for C evidently improves estimation (5), and yields better results than the MSE estimator for FH. The best MSE estimation using (6) is obtained for composition opt. Composite estimator C only approximates the optimal one and, therefore, its MSE estimator makes larger errors. On the other hand, these errors are acceptable if to compare them with the results for FH.

Table 1: Average RMSEs and ABs of estimators for the unemployed fractions.

Estimator	Average RMSE ($\times 10^2$)				Average AB ($\times 10^2$)			
	Domain size class				Domain size class			
	any	small	med.	large	any	small	med.	large
$\hat{\theta}_i^{\mathrm{d}}$	2.4793	3.8540	2.4578	1.1259	0.0636	0.1200	0.0485	0.0223
$\hat{\theta}_i^{\mathrm{S}}$	1.8174	2.8950	1.5632	0.9940	1.3461	2.3656	1.0677	0.6050
$\hat{\theta}_i^{\mathrm{FH}}$	1.7857	2.6707	1.7156	0.9707	0.7349	1.4738	0.5496	0.1811
$\hat{\theta}_i^{\mathrm{C}}$	1.7511	2.6798	1.6838	0.8897	0.7951	1.4777	0.6130	0.2946
$\hat{\theta}_i^{\mathrm{opt}}$	1.4712	2.3804	1.2486	0.7846	0.7301	1.3978	0.5206	0.2720
$\mathrm{mse}(\hat{\theta}_i^{\mathrm{FH}})$	0.0223	0.0445	0.0173	0.0051	0.0180	0.0373	0.0128	0.0039
$\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{C}})$	0.0708	0.1540	0.0491	0.0094	0.0263	0.0532	0.0215	0.0041
$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}})$	0.0173	0.0371	0.0119	0.0030	0.0135	0.0296	0.0087	0.0021
$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{opt}})$	0.0098	0.0206	0.0064	0.0023	0.0050	0.0110	0.0027	0.0012

5 Conclusions

The construction of composite estimator (8) is based on the monotonicity of the variance of the direct estimator as the function of the proportion. Approximation (7) is the monotone function in two separate parts of the interval $[0, 1]$. Therefore, the composition loses its efficiency for the proportions close to turning point $1/2$.

The simulations show that the new composition might be an alternative to the classical EBLUP estimating small proportions in small domains. Design-based estimators and estimators of MSE under the design-based approach are desirable in practice. That design MSE estimator (6) works well in the experiment.

References

Čiginas, A. (2021). Design-based composite estimation rediscovered. In *Proceedings of the 63rd ISI World Statistics Congress*, The Hague, The Netherlands. (to appear).

Fay, R. and Herriot, R. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366):269–277.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11(1):55–77.

Rao, J. and Molina, I. (2015). *Small Area Estimation*. John Wiley, New Jersey, 2 edition.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

SAE estimation under coherence for different overlapping areas. An application for the estimation of employment and unemployment from LFS for Cities and Fua

Michele D'Aló [a*], Danila Filipponi [b*] and Silvia Loriga [c*]

[*]Istat, Italy

Abstract

The aim of this work is to describe the statistical methodology used to produce estimates of a set of labour market variables at City and Fua level and to analyse the results obtained. The small area estimates have been computed through a unit level multivariate model, designed to allow the estimation of the variables of interest in a coherent way and to exploit the administrative data from the Integrated System of Registers (SIR). The estimator is based on a multivariate model implemented through the R MIND function, developed by Istat. The method described in the present work is an extended multivariate version of the more standard linear mixed model at unit level. The coherence of indicators across different domains was reached via a single cross-classification model that includes all the domains of interest. The results show significant efficiency's gains with respect to the direct estimates, this is relevant in particular for the estimation of the unemployed persons (total and by sex) for which the sampling errors of direct estimates were rather high.

Keywords— multivariate small area models, registers and administrative data, coherence among estimates

1 Goal of the work

The Labour force survey (LFS) is the main source of information on the Italian labour market, more specifically on the labour supply, that is employment and unemployment of the population. Beyond the direct estimates computed for

[a] dalo@istat.it
[b] dafilipp@istat.it
[c] siloriga@istat.it

the usual administrative domains (representing planned domains in the LFS sampling design), small area estimates are regularly produced for the Labour Market Areas, that is sub-regional areas where the bulk of the labour force lives and works, and where companies can find the largest amount of the needed labour force. In recent years the interest in statistical information on small geographical areas increased; Eurostat (EUROSTAT, 2018) in agreement with National Statistical Institutes is promoting the production of estimates on Cities and their commuting zones (the so-called Functional Urban Areas – Fua).
Fua are based on the OECD-EC city definition and they represent territories that are highly integrated from an economic point of view. Such territorial domains are typically "small areas": they are not planned domains of social surveys, some of them result uncovered by the sample, in general sampling rate is very variable; they often intersect the administrative boundaries of provinces (Nuts3) and, in some cases, even of regions (Nuts2).
Regarding the labour market participation, the variables of interest on such areas are twelve, namely economically active population (EAP), total and age class 20-64, by gender; employed (EMP) age class 20-64, by gender; unemployed total, by gender. Beyond the additive coherence between estimates referred to men, women and total, it is worth noting that a relationship holds between the indicators:

$$EAP\_20-64 < EMP\_20-64 + UNE < EAP \quad sex = M, F, Total.$$

In order to produce small area estimates, the availability of covariates is fundamental. We dispose of a large set of administrative information for all the individuals in the population deriving from the Integrated System of Registers, in particular from the Thematic Labour Registry (RTL) and the Base Register of Individuals (RBI), integrated with further information on demographic and social aspects, mainly employment benefits and income. The aim of this work is to describe the statistical methodology to produce estimates of some labour market variables at City and Fua level using small area models, exploiting administrative data beyond LFS data and guaranteeing the coherence among all the estimates. In paragraph 2 the small area estimation models that we experimented are described. Main results are analysed in paragraph 3, comparing small area estimates obtained through different models with direct ones; finally, some conclusions are drawn.

2 Small area estimation models

As we mentioned earlier Cities and Fua do not represent planned domains of the LFS. The survey's planned domains are provinces and regions, for which annual and quarterly direct estimates are released. The sample design envisages a stratification of municipalities at provincial level and the calibration used to produce the direct estimates is carried out with respect to the population of provinces and regions. Since Cities and Fua are not planned domains, the sample coverage varies and certain areas may be uncovered.

The direct LFS estimates of the twelve indicators mentioned earlier were produced for the areas of interest (year 2018); they represent the first step towards the production of estimates with small area models. They were produced applying the same estimator used for the survey's current estimates, namely a calibration estimator for which the constraints are represented essentially by the population distribution by age and sex at different territorial scales (provinces, regions). The accuracy of direct estimates has been evaluated through their coefficient of variation.

Coefficients of variation strongly vary among the areas and the indicators of interest. Looking at two main indicators – employed aged 20-64 (EMP_20-64) and unemployed (UNE), by gender, on Cities and Fua – cvs may be considered "acceptable" for the employed (the maximum is 0.22), while for the unemployed they are higher: they exceed 0.33 in 29 Fua and 58 Cities – maximum values are 0.57 in one Fua and 1.10 in one City –, while they range between 0.16 and 0.33 in 114 Fua and 136 Cities and in 103 Fua and 64 Cities they are lower than 0.16. The small area estimation model for achieving the estimation goal of the present work must follow a multivariate approach in order to preserve the inner coherence among the estimates of the different indicators. The estimator we used is based on a multivariate model implemented through the R MIND function (D'Aló et al., 2021a,b), developed by Istat. The method is an extended multivariate version of the more standard linear mixed model at unit level (Battese et al., 1988). The main extensions give the possibility to include two or more random effects in the model and to consider a multivariate qualitative variable as dependent variable (following the multivariate modelling approach of Datta et al., 1999). More specifically, as regards the random effects, when some domains of interest are not covered by the sample, it is possible to introduce one or more marginal random effects that, unlike the domains of interest, are all observed in the sample. In order to produce the labour market estimates at City and Fua level, the potential of the multivariate approach proposed by MIND has been exploited in two ways: to specify the dependent variable and to define the random effects. Since the variables of interest are the employed, the unemployed and the economically active population (that is the sum of employed and unemployed persons), the dependent variable y was defined as a vector composed by three dichotomous variables representing the categories of the employed, unemployed and inactive (these three groups represent an exhaustive and mutually exclusive classification of the population). Obviously, the joint modelling of the three labour market status guarantees the coherence with the population of the domain of interest. The territorial domains of interest for each variable are given by Cities and Fua. Within each territorial domain, the units are further specified according to sex and age group (the indicators for the employed refer to the 20-64 age group, the indicators for the unemployed to those aged 15 and over, while the indicators for the economically active population refer to both age groups).

In order to guarantee the coherence of indicators across different domains, it is possible to define a single cross-classification model that includes all the domains of interest. The model for the vector $y_{d,j,k}$ associated to the k ($k = 1, \dots, N_{d,j}$)

individual in the domain (d, j), being $d = 1, \ldots, D$ the geographical domain and $j = 1, \ldots, C$ the modalities of socio-demographic variables, can be expressed as:

$$y_{d,j,k} = \mathbf{X}_{d,j,k}^T \beta + \tau_d + \delta_j + \gamma_{d,j} + e_{d,j,k},$$

where τ, δ and γ are the random effects. We consider a model where the random effects δ and γ are degenerate at zero.
In order to estimate the labour market indicators, the information deriving from the Integrated System of Registers was used and, in particular, that from the Thematic Labour Registry (RTL) and from the Base Register of Individuals (RBI), integrated with further administrative information on demographic and social aspects, mainly employment benefits and income. Besides the usual demographic auxiliary variables, information regarding monthly employment, events of job-protected leave for which the worker receives an allowance, information on the end of the job are used. Moreover, from fiscal sources, information on work income, retirement income and capital income are also available. Exploiting these income variables jointly with the family's characteristics, the individual equivalent income indicator was also computed. Finally, the following social variables are taken into account: educational qualification and participation in training courses; retirement scheme, other types of financial support (unemployment benefits, family allowances for workers, transfers to families with economic problems, sickness and maternity allowances, subsidies for students).

3 Analysis of the results

The selection of the estimation models for the indicators of interest at City and Fua level was carried out considering separately the employed and the unemployed. The best fit of the model has been assessed through a stepwise approach applied to the full mixed model, defined for Fua and City, maximising the usual diagnostics such as the AIC and BIC criteria (see Table 1). The marginal and conditional R^2 have been computed; as expected, the marginal R^2 is very high (77%) for the employed, thanks to the presence of variables that are strongly associated to the employment status such as administrative employment and work income. The marginal R^2 is instead lower (14%) for the unemployed, as there are no variables strongly associated to the phenomenon. The random domain effect leads to a small increase of the R^2 index (conditional R^2).

Table 1: Indicators of the goodness of fit of the selected model

area	Indicator	N_Obs	AIC	BIC	LL	LLDF	Sigma	Marginal R^2	Conditional R^2
FUA	Employed	231504	-17142.4	-16697.2	8614.20	43	0.233	0.773	0.775
FUA	Unemployed	231504	-76704.8	-76259.7	38395.42	43	0.205	0.132	0.137
City	Employed	144899	-9689.8	-9264.8	4887.92	43	0.233	0.771	0.773
City	Unemployed	144899	-40462.9	-40037.9	20274.49	43	0.210	0.136	0.142

That is confirmed also by the analysis of the inter-class correlation coefficient (ICC). For both the employed and the unemployed, as well as for both the domains (Fua and Cities), the variance is almost fully explained by the variability across units, while the variability explained by the inclusion in the different domains is almost equal to zero. Based on these results we can expect that the efficiency and reliability of small area estimates will strongly depend on the auxiliary information used in the estimation model and in particular on their relationship with the variable of interest. To this aim, the availability of a large set of administrative information is very useful in the estimation process. Looking at the distribution of the small area estimates and the corresponding direct estimates, taking into account also their confidence intervals, we see that the SAE estimates do not present evident systematic bias (see Fig. 1). In particular, we may see that the small area estimates fall almost entirely in the confidence interval of the direct estimates. This applies in general to the whole set of parameters of interest as well as to both Cities and Fua. In some cases, outliers of the direct estimates were corrected by the SAE estimates.

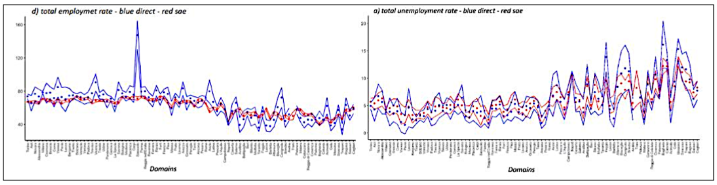


Figure 1: SAE and direct estimates for employment and unemployment rate and related confidence intervals. Domains: Cities

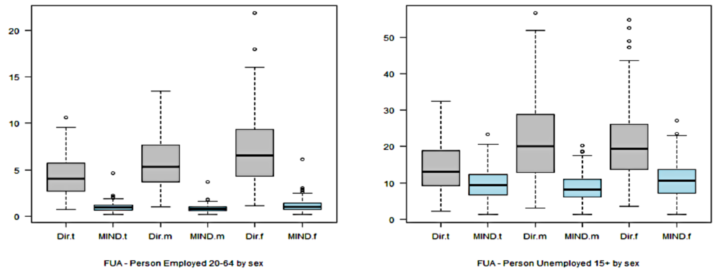


Figure 2: Distribution of coefficient of variations (percentage) of SAE and direct estimates for employed and unemployed by gender. Domains: Fua

In general, the SAE estimates lead to strong efficiency gains for the employed, due to the goodness of the fitted model, but also for the unemployed the gain of efficiency is relevant (see Fig. 2).

4 Conclusions and further developments

In this work we described the methodology used to produce SAE estimates for several labour market indicators over different domains (Fua and Cities), which are not planned domains for the LFS and intersect administrative units (provinces and regions). We adopted a unit level multivariate model, designed to allow the estimation of the variables of interest in a coherent way and to exploit the large set of administrative information from the Integrated System of Registers (SIR). Beyond this estimator, area level univariate models have also been tested, however the use of a specific model for each variable of interest does not guarantee the inner coherence among the estimates; moreover the informative potential from the SIR is better exploited by the unit level model, while area level estimates tends to be closer to direct ones; compared with direct estimates, SAE unit level models allow to obtain more evident efficiency gains. These estimates are made available online on Eurostat website[1] as part of the Cities' database and they enlarge the set of information on Italian labour market over small territorial areas. Ongoing studies concern the analysis of the external coherence of these SAE estimates with direct LFS estimates usually produced and disseminated. Benchmarking techniques should also be taken into account, considering that LFS direct estimates are produced and disseminated at several levels of disaggregation and, in a cross-sectional perspective, LFS provides monthly estimates for the whole Country, quarterly figures for NUT2 regions and yearly figures at NUTS3 level.

References

Battese G. E., Harter R. M. and Fuller W. A. (1988). An Error-Components Model for Prediction of County Crop Areas Suing Survey and Satellite Data. *Journal of the American Statistical Association* 83(401), 28–36

D'Aló, M., Falorsi, S. and Fasulo (2021). Multivariate InfereNce for Domains - MIND methodology and R Package. *SAE2021 Conference*

D'Aló, M., Falorsi, S. and Fasulo (2021). MIND R Package *https://cran.r-project.org/web/packages/mind/index.html*

Datta G. S., Day B., Basawa I. (1999). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference*, 75(2), 269–279.

EUROSTAT (2018). Methodological manual of territorial typologies. *https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Territorial_typologies_manual*.

[1] EUROSTAT Cities database: ,https://ec.europa.eu/eurostat/web/cities/data/database

MIND, an R package for multivariate small area estimation with multiple random effects.

Michele D'Alò [a*], Stefano Falorsi [b*] and Andrea Fasulo [c*]

[*]Istat, Italy

Abstract

This work is aimed to describe a Small Area Estimator based on a multivariate linear mixed model implemented in the MIND R Package (2021a). The method is a multivariate version of the standard area and unit small area estimators based on linear mixed model. Beyond the possibility of considering multivariate qualitative (or quantitative variable in the incoming release of the new version of package) dependent variable, the proposed method allows to specify a model with more than one random effects. The further marginal random effects, in addition to the usual random area effects, can be very useful when the areas of interest are very small and a significant numbers of areas are out of sample. This means that when some domains of interest are out of sample, or are strongly under covered, it is possible to introduce one or more marginal random effects that are instead well represented in the sample data. In this way, the bias of the synthetic estimator could be lessened. The estimates obtained with MIND fall within the group of estimators classified as model-based projection or composed estimator. The last one using multivariate predicted values, under the model, only for the subset of units that are not included in the sample. The REML estimation of variance components of the mixed effects model and MSE is calculated by a multivariate extension of the methodology proposed in Saei e Chambers (2003). A simulation study has been carried out to evaluate the performance of the small area estimators considered, while an application to real data are report in D'Alò et al (2021b).

Keywords— ANOVA, Spatial and Temporal Autocorrelation, EBLUP

1 Introduction

Over the last few years, the paradigm underlying the statistical process has been gradually changing the production of official statistical data by the

[a] dalo@istat.it

[b] stfalors@istat.it

[c] fasulo@istat.it

major information statistical centres, both nationally and internationally. In fact, alongside the data collected using traditional "statistical surveys", the growing availability of data from so-called "new data sources" – both those of an administrative nature and those obtained through new electronic devices and information-gathering channels on the Internet – overwhelmingly dictates the agenda of the methodological and operational aspects to be addressed and resolved by official statisticians in each country. As far as the more strictly statistical-methodological aspects are concerned, the following aspects are relevant: the need to estimate multiple contingency tables, which arises from the fact that large scale surveys produce multiway tables (hypercubes) obtained from the intersection of numerous variables; many different and intersecting territorial and structural estimation domains; the different cells of hypercubes may be either estimated using information arising from Statistical Registers or estimated using survey data by means of direct or indirect estimators; the need to produce predicted values at the level of each the single record of the Statistical Register representing the target population. In this context, multivariate modelling may be more efficient (or appropriate) because there are multiple target variables from each small area, and these are either correlated with (or mutually restrictive of) each other. Also, a set of counts may sum up to a known total in each small area, such as the number of persons in different household types or the number of persons with the three different labour market statuses. The small area estimation methodology proposed, based on multivariate small area model, can properly aid to the solution of the above listed problems. For estimating simultaneously different totals of interest on the basis of the same multivariate linear mixed model, the multivariate modelling approach by Datta et al (1999) is followed. In particular, the proposed Small Area Estimation (SAE) method is a multivariate extension of the standard estimators for small area based on a linear mixed model with only an area random effects, in fact, it allows to deal with more than one target variable at same time. From this the name of the R package MIND - Multivariate model based INference for Domains (`https://cran.r-project.org/web/packages/mind/index.html`). Moreover, the method here described allows a model specification with more than one random effect, so that, possible marginal effects can be fitted in the model. This further possible marginal random effects, in addition or instead to the usual random area effects, can be very useful when the areas of interest are very small and a significant numbers of areas are out of sample. This means that when some domains of interest are out of sample or are strongly under covered, one or more marginal random effects, which are covered in the sample data, can be fitted into the model. Bigger is the deviance among marginal effects more the bias of the synthetic estimator can be lessened. This means that more local smoothed synthetic estimates can be derived for the out of sample areas, and that generally the over-shrinkage of the small area estimates can be smaller. The marginal random effect may be derived from the variables used to define the strata or from some other variables utilized for defining the planned domains or for cross-classify the population units.

2 MIND methodology

Let's consider the General Linear Mixed Model (GLMM) with: $V \geq 1$ target variables, $\Delta \geq 1$ and $J \geq 1$ factors for the fixed and random part of the model. The v-th ($v = 1, \dots, V$) target variable includes $V_k \geq 2$ categories or $V_k = 1$ for a quantitative variable. The δ-th ($\delta = 1, \dots, \Delta$) fixed effect factor has $G_\delta \geq 2$ levels, where g_δ denotes the generic of them ($g_\delta = 1, \dots, G_\delta$); $G_\delta = 1$ for a quantitative variable. The j-th ($j = 1, \dots, J$) random effect factor is characterized by Q_j levels, where q_j denotes the generic of them ($q_j = 1, \dots, Q_j$). Then the multivariate model considers $V (= \sum_{k=1}^{K} V_k)$ target variables, $G (= \sum_{\delta=1}^{\Delta} G_\delta)$ levels for the fixed part and $Q (= \sum_{j=1}^{J} Q_j)$ for the random part.
The GLMM, $M(\mathbf{y};\ \boldsymbol{\beta},\ \boldsymbol{\omega})$, depends on a set of unknown parameters: regression coefficients $\boldsymbol{\beta}$ and variance components $\boldsymbol{\omega}$ and is expressed as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$. In order to deal with the multivariate nature of the model, let's denote with: $* = [*_b]$, a column vector, $* \equiv \mathbf{a}$, or a matrix, $* \equiv \mathbf{A}$, formed by B blocks: $*_b \equiv \mathbf{a}_b$ or $*_b \equiv \mathbf{A}_b$, respectively; $diag_{b=1}^{B} *_b$ a diagonal block matrix of B blocks $*_b$ (scalar a_b, vector $\mathbf{a}_b$ or matrix $\mathbf{A}_b$); $\boldsymbol{\Sigma}_{\mathbf{a}} = diag_{b=1}^{B} \mathbf{a}_b$ a diagonal block matrix formed from vector $\mathbf{a}$; $mat_{b,b'}^{B} \{*_{bb'}\}$ a block matrix formed by $B \times B$ blocks: $*_{bb'} \equiv \mathbf{a}_{bb'}$ or $*_{bb'} \equiv \mathbf{A}_{bb'}$; $\otimes$ the Kronecker product. The matrices and vectors of the model, $\mathbf{y} = [\mathbf{y}_i]$, $\mathbf{X} = [\mathbf{X}_i]$, $\boldsymbol{\beta} = [\boldsymbol{\beta}_{\delta, g_\delta}]$, $\mathbf{Z} = [\mathbf{Z}_i], \mathbf{u} = [\mathbf{u}_{j, q_j}]$ and $\mathbf{e} = [\mathbf{e}_i]$, are formed by column blocks each one composed by V rows. The index $i (i = 1, \dots, a)$ denotes the generic basic element of the model: sampled unit or area (domain). More specifically, $\mathbf{y}_i$ and $\mathbf{e}_i$ denote vectors of target variables and of residuals, $\mathbf{X}_i {=}\ \dot{\mathbf{x}}_i' \otimes \mathbf{I}_V$ and $\mathbf{Z}_i {=}\ \dot{\mathbf{z}}_i' \otimes \mathbf{I}_V$ are the design matrices of fixed and random effects - being $\mathbf{x}_i'$ and $\mathbf{z}_i'$ two covariate vectors available for the i-th basic element of the model; $\boldsymbol{\beta}_{\delta, g_\delta}$ and $\mathbf{u}_{j, q_j}$ are vectors referred to g_δ-th and q_j–th levels of δ-th fixed and j-th and random effects factors respectively. Notice that $\dot{\mathbf{x}}_i'$ may include a subset of $\Delta' \leq \Delta$ quantitative variables, being for these $G_{\delta'} = 1 (\delta' = 1, \dots, \Delta')$.
Vector $\mathbf{e}$ is supposed to have an *iid* multi-Normal, $\mathbf{MN}(\mathbf{0}, \mathbf{R})$, distribution with mean $\mathbf{0}$ and variance covariance matrix $\mathbf{R} = \mathbf{R}(\boldsymbol{\sigma}_e^2)$ with a diagonal block structure dependent on the vector of variance components $\boldsymbol{\sigma}_e^2 = [\sigma_{e,v}^2]\quad (v = 1, \dots, V)$. In particular, $\mathbf{R} = [\boldsymbol{\Sigma}_{\boldsymbol{\sigma}_e} \otimes \mathbf{I}_a] \bullet \mathbf{W}$ with $\boldsymbol{\sigma}_e = [\sigma_{e,v}^2]$ and $\mathbf{W} = diag_{i=1}^{a} \mathbf{W}_i$ and $\mathbf{W}_i = \grave{w}_i \otimes \mathbf{I}_V$ for $\grave{w}_i$ a known quantity assigned to i-th basic element of (1).
Vector $\mathbf{u}$, is supposed to have an *iid* multi-Normal, $\mathbf{MN}(\mathbf{0}, \mathbf{G})$, distribution with mean $\mathbf{0}$ and a block diagonal structure for variance covariance matrix $\mathbf{G} = \mathbf{G}(\boldsymbol{\omega})$ dependent on the vector of $(1 + 2 \bullet J) \times V$ variance components $\boldsymbol{\omega} = [\boldsymbol{\sigma}_e^{2'}, \gamma']'$, being $\boldsymbol{\gamma} = [\boldsymbol{\varphi}\prime, \boldsymbol{\rho}\prime]'$ for $\boldsymbol{\varphi} = [\boldsymbol{\varphi}_j]$ and $\boldsymbol{\rho} = [\boldsymbol{\rho}_j]$. In details, $\mathbf{G} = [\boldsymbol{\Sigma}_{\boldsymbol{\sigma}_e} \otimes \mathbf{I}_Q] \bullet \boldsymbol{\Omega}$, where $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\gamma}) = diag_{j=1}^{J} \boldsymbol{\Omega}_j$ is defined under a linear covariance structure composed by J diagonal blocks. The structure of j-th block $\boldsymbol{\Omega}_j = \boldsymbol{\Omega}_j(\boldsymbol{\varphi}_j, \boldsymbol{\rho}_j)$ is known unless the variance component vectors $\boldsymbol{\varphi}_j = [\varphi_{j,v} = {\scriptstyle (\sigma_{j,v}^2 / \sigma_e^2)}]$ and $\boldsymbol{\rho}_j = [\rho_{j,v}]$. In particular it is obtained as $\boldsymbol{\Omega}_j = \left[\boldsymbol{\Sigma}_{\boldsymbol{\varphi}_j} \otimes \mathbf{I}_{Q_j}\right] \bullet \grave{\boldsymbol{\Omega}}_j$, where

$\boldsymbol{\Sigma}_{\boldsymbol{\varphi}_j} = diag_{v=1}^{V}\varphi_{j,v}$ and $\grave{\boldsymbol{\Omega}}_j = \grave{\boldsymbol{\Omega}}_j(\boldsymbol{\rho}_j)$ is a square matrix formed by blocks, $\grave{\boldsymbol{\Omega}}_{j,q_j q_j'} = \grave{\boldsymbol{\Omega}}_{j,q_j q_j'}(\boldsymbol{\rho}_j)$ $(q_j,\ q_j' = 1,\ldots,Q_j)$, each one of order V. As it is assumed that the variables y_v and $y_{v\prime}(v \neq v\prime)$ are uncorrelated, each one of the $Q_j \times Q_j$ blocks, $\grave{\boldsymbol{\Omega}}_{j,q_j q_j'}$, of $\grave{\boldsymbol{\Omega}}_j = mat_{q_j q_j'}^{Q_j}\grave{\boldsymbol{\Omega}}_{j,q_j q_j'}$ is a diagonal matrix. For each one of the matrices, $\grave{\boldsymbol{\Omega}}_j\ (j = 1,\ \ldots, J)$, we assume three alternative models $M_1 \div M_3$. The first one is the basic ANOVA model, in which $\boldsymbol{\rho}_j \equiv \mathbf{0}$ and: $\grave{\boldsymbol{\Omega}}_{j,q_j q_j'} \equiv \mathbf{0}_{V\times V} for q_j \neq\ q_j'\ ;\grave{\boldsymbol{\Omega}}_{j,q_j q_j'} \equiv \mathbf{I}_V for q_j =\ q_j'$. M_2, assumes an AR(1) process, in which the difference function, $f_j(q_j, q_j^{'})$, between levels q_j and q_j' of j-th random factor, coincides with the lag, $l_{(q_j,\ \ q_j')}$ and $\grave{\boldsymbol{\Omega}}_{j,q_j q_j'} \equiv [I_V(I_V - \boldsymbol{\Sigma}_{\rho j}^2)]^{-1}\boldsymbol{\Sigma}_{\rho j}^{l}(qj,qj')$. M_3, assumes a process depending on a spatial distance function, $s_{(q_j,q_j')}$, and $\grave{\boldsymbol{\Omega}}_{j,q_j q_j'} \equiv \mathbf{I}_V + \delta_{q_j,q_j'} + exp^{s_{(q_j,q_j')}}\boldsymbol{\Sigma}_{\rho_j}^{\mathbf{-1}^{\mathbf{-1}}}$. The BLUE estimator of fixed effects, $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}\,(\boldsymbol{\gamma})$, and BLUP estimator the random effects $\widetilde{\mathbf{u}} = \widetilde{\mathbf{u}}\,(\boldsymbol{\gamma})$ depends on the unknown variance components $\boldsymbol{\gamma}$. Adapting the REML procedure by Saei e Chambers (2003) at the multivariate context, we get $\boldsymbol{\omega} = \hat{\boldsymbol{\omega}}$, and the final plug-in estimators, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}\,(\hat{\boldsymbol{\gamma}})$ and $\hat{\mathbf{u}} = \hat{\mathbf{u}}\,(\boldsymbol{\gamma})$, are expressed as: $\hat{\boldsymbol{\beta}} = \hat{\mathbf{M}}_{XX}^{\mathbf{-1}}\,(\hat{\gamma})\bullet\hat{\mathbf{m}}_{Xy}\,(\hat{\gamma})$ and $\hat{\mathbf{u}} = \hat{\mathbf{T}}^{*}[\hat{\mathbf{m}}_{Zy}\,(\hat{\gamma}) - \hat{\mathbf{M}}_{ZX}\,(\hat{\gamma})]$. In the above formulas: $\hat{\mathbf{M}}_{XX}^{\mathbf{-1}}\,(\hat{\gamma}) = [MXX - MZX'T*MZX]-1$, $\hat{\mathbf{m}}_{Xy}\,(\hat{\gamma}) = [\mathbf{m}_{\mathbf{Xy}} - \mathbf{M}_{ZX}'\hat{\mathbf{T}}^{*}\ \mathbf{m}_{Zy}]$ in which $\hat{\mathbf{T}}^{*} = \hat{\mathbf{T}}^{*}\,(\hat{\gamma})$ is given by $\hat{\mathbf{T}}^{*} = [\mathbf{M}_{ZZ} + \boldsymbol{\Omega}^{-1}(\hat{\gamma})]^{-1}$, where $\mathbf{M}_{AB} = \mathbf{A}'\mathbf{W}\mathbf{B}$ and $\mathbf{m}_{Ab} = \mathbf{A}'\mathbf{W}\mathbf{y}$ for $\mathbf{A}$ or $\mathbf{B} = (\mathbf{X}, \mathbf{Z})$.

Under GLMM, the general expression of target parameter is given by $\boldsymbol{\eta} = [\boldsymbol{\eta}_d] = [\mathbf{X}_{\mathbf{d}*}^{+}\boldsymbol{\beta} + \mathbf{Z}_{\mathbf{d}*}^{+}\mathbf{u}]$, where $\boldsymbol{\eta}_d$ is the sub-vector of V target parameters for d-th target domain $(d = 1, \ldots, D)$. The EBLUP estimator, $\hat{\boldsymbol{\eta}}^{EP} = \hat{\boldsymbol{\eta}}^{EP}\,(\hat{\gamma})$, of $\boldsymbol{\eta}$ is $\hat{\boldsymbol{\eta}}^{EP} = \left[\hat{\boldsymbol{\eta}}_d^{EP}\right] = [\mathbf{X}_{\mathbf{d}*}^{+}\hat{\boldsymbol{\beta}} + \mathbf{Z}_{\mathbf{d}*}^{+}\mathbf{u}]$. The correspondent Synthetic Predictor is $\hat{\boldsymbol{\eta}}^{SP} = [\mathbf{X}_{\mathbf{d}*}^{+}\hat{\boldsymbol{\beta}}]$. The MSE of $\hat{\boldsymbol{\eta}}^{EP}$ is the diagonal of $MCPE\left[\hat{\boldsymbol{\eta}}^{GE}\right] = \hat{\mathbf{G}}_1\,(\hat{\gamma}) + \hat{\mathbf{G}}_2\,(\hat{\gamma}) + 2\hat{\mathbf{G}}_3\,(\hat{\gamma})$, where: $\hat{\mathbf{G}}_1\,(\hat{\gamma}) = [\boldsymbol{\Sigma}_{\hat{\boldsymbol{\sigma}}_e} \otimes \mathbf{I}_D]\bullet\mathbf{Z}_{*}^{+}\hat{\mathbf{T}}^{*}\mathbf{Z}_{*}^{+'}$; $\hat{\mathbf{G}}_2\,(\hat{\gamma}) = [\boldsymbol{\Sigma}_{\hat{\boldsymbol{\sigma}}_e} \otimes \mathbf{I}_D]\,\hat{\mathbf{G}}_{2*}\ \hat{\mathbf{M}}_{XX}^{\mathbf{-1}}\,(\hat{\gamma})\,\hat{\mathbf{G}}_{2*}'$, for $\hat{\mathbf{G}}_{2*}\,(\hat{\gamma}) = \left[\mathbf{X}_{*}^{+} - \mathbf{Z}_{*}^{+}\hat{\mathbf{T}}^{*}\,(\hat{\gamma})\,\hat{\mathbf{M}}_{ZX}\,(\hat{\gamma})\right]$; $\hat{\mathbf{G}}_3\,(\hat{\gamma}) = [\boldsymbol{\Sigma}_{\hat{\boldsymbol{\sigma}}_e} \otimes \mathbf{I}_D]\,tr\left\{\hat{\mathbf{C}}\ \hat{\boldsymbol{\Sigma}}^{*}\hat{\mathbf{C}}'\hat{\mathbf{B}}\right\}$. In last formula: $\mathbf{C} = \partial[\mathbf{Z}_{*}^{+}\mathbf{T}^{*}\,(\gamma)/\partial\boldsymbol{\gamma}]$, which equals to $-\mathbf{Z}_{*}^{+}\mathbf{T}^{*}\,(\gamma) \otimes \mathbf{I}_{2\times J}/\partial\boldsymbol{\Omega}^{-1}\,(\gamma)\partial\boldsymbol{\gamma}\mathbf{T}^{*}\,(\gamma)$; $\boldsymbol{\Sigma}^{*} = \mathbf{M}_{\mathbf{ZZ}} + \mathbf{M}_{\mathbf{ZZ}}\boldsymbol{\Omega}(\gamma)\mathbf{M}_{\mathbf{ZZ}}$; $\mathbf{B}$ is the submatrix of the inverse of the Fisher information matrix referred to the sub-component $\boldsymbol{\gamma}$ of the vector of the variance components $\boldsymbol{\omega}$.

The setting of $\hat{\boldsymbol{\eta}}^{GE}$ is completely general and allows to apply with the same general formulation both unit or area level estimators, depending on the level of data aggregation of the basic elements of the model. In case of a unit-level estimator, the basic elements of the model coincide with the sampling units, then: $a \equiv n$ and $\equiv k \quad (k = 1, \ldots, n)$. In case of an area-level estimator, the model coincide with the sampled domains, then: $a \equiv \grave{D}$ and $\equiv d \quad (d = 1, \ldots, \grave{D})$, being D, the size of the population domains and $D_{out} = D - \grave{D}$ the out of sample ones.

3 Simulation study

In order to evaluate the performance of the proposed estimator a Monte Carlo simulation 200 samples have been drawn from the 2011 Italian Population Census for one Italian region, Piedmont, using the Italian Permanent Census sampling design. In this region, there are 1201 municipalities and 359 of those are included in the sample. The target variables are the population counts for the five modes of the variable occupational status (employed, unemployed, retired, student, in other condition). The overall number of cells in the census table is 12010 (1201 municipalities times 2 gender times 5 occupational status). A mixed linear model with two random effect has been specified, the first at municipality level by gender and the second at level of aggregated Local Labour Market Area. The auxiliary variables are demographic variables, i.e. class of age by gender, marital status, educational level and citizenship. Different estimators are compared by means of the standard indicators of accuracy of prediction: Average Absolute Relative Bias (AARB) and Average Relative Root Mean Squared Error (ARRMSE). The evaluation indicators are formulated as follows:

$$AARB = \frac{1}{D}\sum_{d=1}^{D}\frac{\left|\frac{1}{R}\sum_{r=1}^{200}\hat{y}_{rd} - Y_d\right|}{Y_d} \qquad RRMSE = \frac{1}{D}\sum_{d=1}^{D}\frac{\frac{1}{R}\sum_{r=1}^{200}\sqrt{(\hat{y}_{rd} - Y_d)^2}}{Y_d}$$

where $\hat{y}_{rd}$ and Y_d are the predicted true value in the r-th simulated sample in the domain d. The estimator consider are direct estimator, EBLUP, Projection and Synthetic and their performance in term of AARB and ARRMSE are showed the Table 1 and 2. The simulation study shows good performance of EBLUP and Projection estimator in term of ARRMSE and, except for unemployed and retired people, also in term of AARB, with respect to the direct estimator for in sample areas.

4 Conclusions and further developments

The `R` package MIND allows to compute small area estimates based on multivariate mixed model using two and more random effects. The simulation study shows good performance of proposed estimators. Further development will be the introduction of the spatial correlation among areas and the time correlation for repeated surveys.

References

D'Alò M., Falorsi S. and Fasulo A. (2021) . MIND R Package `https://cran.r-project.org/web/packages/mind/index.html`

Table 1: AARB for the variable occupational status for the estimators

Direct	**Employed**	**Unemployed**	**Student**	**Retired**	**Other**
In-sample	7,2	7,4	7,8	7,3	7,4
EBLUP	**Employed**	**Unemployed**	**Student**	**Retired**	**Other**
Overall	4,0	34,2	5,1	14,5	7,7
In-sample	1,8	16,8	2,1	6,8	3,2
Out-sample	4,9	41,6	6,4	17,8	9,6
Projection	**Employed**	**Unemployed**	**Student**	**Retired**	**Other**
Overall	4,3	37,7	5,5	16,2	8,4
In-sample	3,2	28,6	3,4	12,6	5,5
Out-sample	4,9	41,6	6,4	17,8	9,6
Synthetic	**Employed**	**Unemployed**	**Student**	**Retired**	**Other**
Overall	5,0	41,1	6,3	16,8	9,5
In-sample	4,7	37,7	5,8	14,7	8,9
Out-sample	5,1	42,6	6,5	17,8	9,7

Table 2: ARRMSE for the variable occupational status for the estimators

Direct	**Employed**	**Unemployed**	**Student**	**Retired**	**Other**
In-sample	12,3	38,5	14,4	31,5	17,3
EBLUP	**Employed**	**Unemployed**	**Student**	**Retired**	**Other**
Overall	4,6	27,3	5,6	14,1	8,7
In-sample	3,5	22,2	4,0	11,3	6,0
Out-sample	5,1	29,5	6,3	15,3	9,8
Projection	**Employed**	**Unemployed**	**Student**	**Retired**	**Other**
Overall	4,7	27,8	5,7	14,4	8,9
In-sample	3,9	23,8	4,4	12,2	6,8
Out-sample	5,1	29,5	6,3	15,3	9,8
Synthetic	**Employed**	**Unemployed**	**Student**	**Retired**	**Other**
Overall	5,2	28,6	6,2	14,6	9,3
In-sample	4,9	26,7	5,8	13,0	8,7
Out-sample	5,3	29,4	6,4	15,3	9,6

D’Aló, M., Filipponi D. and Loriga, S. (2021). SAE estimation under coherence for different overlapping areas. An application for the estimation of employment and unemployment from LFS for Cities and Fua. *SAE2021 Conference*

Datta G. S., Day B., Basawa I. (1999). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference*, 75(2), 269–279.

Saei, A. and Chambers, R., (2003). ’Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects. *S3RI Methodology Working Paper M03/15.*

Estimation of the Employment Rate by Municipality. Interpreting Results from the SAE model

Enrique de Alba [a*] and Eric Rodríguez [b*]

[*]Instituto Nacional de Estadística, Geografía e Informática, Mexico

Abstract

In the first quarter of 2019 there were 2457 municipalities in Mexico, distributed among 32 states. The size of the total population in each one is very variable. The working age population (older than 15 years, according to Mexican legislation) in these municipalities goes from 68 people in Santa Magdalena Jicotitlán, Oaxaca to 1,448,788 in Iztapalapa, Mexico City. In this context, the analysis of results from applying small area estimation models to estimate employed population in each one of these municipalities requires identifying auxiliary variables that provide sensible results. In a study of the employment rate by means of estimating both employed and total economically active populations by municipality, the resulting model shows an employment rate equal to, or larger than 99 percent of the economically active population for 328 municipalities, whose population older than 15 varies from 167 to 85,875 persons. Economic theory about the Natural unemployment rate indicates that in any economy there exists a subset of the total economically active population, which is part of frictional unemployment, defined as those who are moving from one job to another or that just entered the labour market. This makes it very unlikely to have employment rates of 100 percent, where all people interested in getting a job have one, since this would motivate part of those population in the economically unactive who are available to work, who could try to enter in labour market. In this paper we look for variables that allow us to identify results from the model that are adequate to show the reality of the employment rate at the municipality level.

Keywords— Small area, unemployment, municipality, Mexico

[a]enrique.dealba@inegi.org.mx
[b]eric.rodriguez@inegi.org.mx

1 Introduction

Article 3 of the National Statistics and Geographical Information System of Mexico (SNIEG, in Spanish) establishes that the System has as its aim to provide the State and society as a whole with Information that is of quality, relevant, timely and truthful in order to contribute to national development. The guiding principles of the System will be accessibility, transparency, objectivity and independence. The Mexican Government is divided into three orders: Federal, State and Municipal. The three orders contribute to the attainment of the State goals and tasks: guarantee human rights, provide public services, assure law and order, write new laws and monitor their implementation. So whenever possible it is necessary to produce statistical information disaggregated by all orders of government.

2 Background

The National Survey of Occupation and Employment (ENOE) is the main source of information on the Mexican labour market; it provides monthly and quarterly data on the labor force, employment, labor informality, underemployment and unemployment. It is also the largest continuous statistical project in Mexico, providing national and four-size locality figures for each of the 32 states and for a total of 39 cities.
The general objectives of the ENOE are:

1. Ensure that the country has basic statistical information with national, state and major city representation on the employment characteristics of the population.

2. To provide sociodemographic statistical information to complement and deepen the analysis of the employment characteristics of the Mexican population.

3. Increase the supply of strategic indicators for the full knowledge of the national reality and for decision-making in the formulation of labor policies.

The information generated by the ENOE is important for the design of public policies on employment, and data are required with representation at the level of the municipalities. Given these circumstances and the high cost of generating samples that are sufficiently large to calculate indicators at the municipal level. It was necessary to find methodological tools that would allow us to provide data on the participation of the population in the labor market within municipalities: as a result, small area estimation techniques were applied to obtain the number of economically active and employed population. (Vielma Orozco et al., 2021).

3 Small Area Estimation

The purpose of applying SAE is to obtain estimates of the economically active population and the employed (and unemployed) population by municipality, for which census data, surveys and administrative records were explored.

3.1 The Model

An initial set of 12 auxiliary variables was considered – for which temporal and geographic reference, adjustments were made to make them compatible with the variables provided by ENOE, Vielma et al. (2021). The final selection of auxiliary variables reduced to three after testing for their significance:

- Economic dependency ratio (Population under 15 years old and 65 years old and over compared to the population of 15 to 64 years old)
- Proportion of male population (Male population aged 15 to 44 years old compared to population aged 15 and over)
- Proportion of the population affiliated to IMSS[1] or ISSSTE[2] (Population affiliated to these institutions compared to the population aged 15 years old and over).

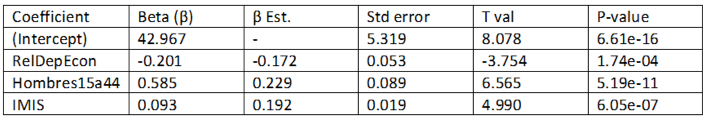
The dependent variables (one model for each) are:
Economically Active Population (EAP), composed of people aged 15 and over who had a link with the economic activity or who looked for it in the reference week, so they were employed or unemployed. Mexican law establishes the age of at least 15 years to be able to work.
Employed population (EP), are people aged 15 and over who in the reference week carried out some economic activity for at least one hour. It includes the employed who had a job, but for some reason did not perform it temporarily, without losing the employment link with it; as well as those who helped in some economic activity without receiving a salary. The condition of carrying out an economic activity at least one hour in the reference week was established by the International Labour Organization.
The models fitted were mixed models with a spatial component included in the part that represents the random effects component resulting in a new model called SEBLUP, Vielma Orozco et al. (2021). Model assumptions were checked in each case. The results of estimating the model are given in Table 1 for Economically Active Population (EAP) and in Table 2 for Employed Population (EP).

Moran's Index was used to test spatial correlation; its value was 0.25088 and the corresponding p-value was p=2.2204E-16. The models were estimated with 760 data from municipalities that had data and whose Coefficient of Variation was les tan 20%. (Vielma Orozco et al., 2021).

[1] IMSS: Mexican Social Security Institute.

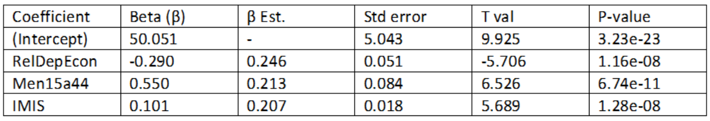
[2] ISSSTE: Institute of Social Security and Services for State Workers.

Table 1: Estimated parameters for the EAP model



Coefficient	Beta (β)	β Est.	Std error	T val	P-value
(Intercept)	42.967	-	5.319	8.078	6.61e-16
RelDepEcon	-0.201	-0.172	0.053	-3.754	1.74e-04
Hombres15a44	0.585	0.229	0.089	6.565	5.19e-11
IMIS	0.093	0.192	0.019	4.990	6.05e-07

Table 2: Estimated parameters for the AP model



Coefficient	Beta (β)	β Est.	Std error	T val	P-value
(Intercept)	50.051	-	5.043	9.925	3.23e-23
RelDepEcon	-0.290	0.246	0.051	-5.706	1.16e-08
Men15a44	0.550	0.213	0.084	6.526	6.74e-11
IMIS	0.101	0.207	0.018	5.689	1.28e-08

3.2 Some Results

In this subsection we comment on some of the results obtained by the SEBLUP models. Figure 1 shows the proportion in the Economically Active population, by municipalities, obtained from the model, INEGI (2019).

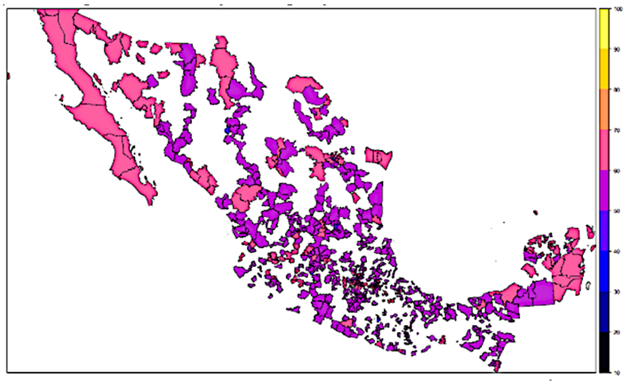


Figure 1: Proportion of EAP by municipality. Source:INEGI. Cifras laborales para los municipios de México, 2019. Estimación para Áreas Pequeñas. Febrary 2020.

The information obtained in the 2020 Population and Housing Census (CPV2020) is a tool to review the results of the SAE model, a linear relationship is observed between the Economically Active Population of both information programs. The same is true of the employed population, so we can consider that the results of the SAE model are adequate. This is shown in Figure 2.

On the other hand, employment rate estimates obtained for all municipalities by combining both models yield the results presented in Figure 3. The largest values are near 100% and many of these are in very small municipalities in the state of Oaxaca. Economic theory about the Natural unemployment rate indicates that in any economy there exists a subset of the total economically

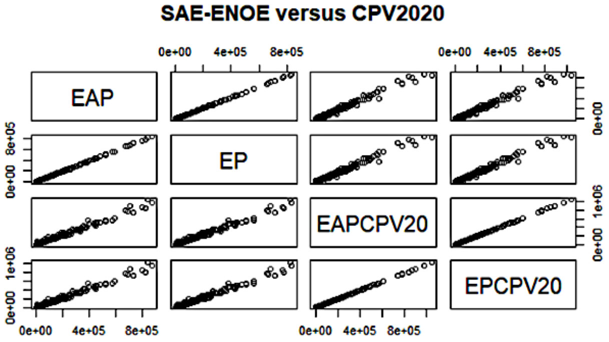


Figure 2: Pairs of the estimates Economically Active Population and the Employed Population from the SAE model and the CPV2020.

active population, which is part of frictional unemployment, defined as those who are moving from one job to another or that just entered the labour market. This makes it very unlikely to have employment rates of 100 percent, where all people interested in getting a job have one, since this would motivate part of those population in the economically unactive who are available to work, who could try to enter in labour market. Thus, further analysis is required to explain this. Perhaps combining these results with those on Informality in Mexico as obtained by Ibarra-Olivo et al (2021) can help understand what is going on.

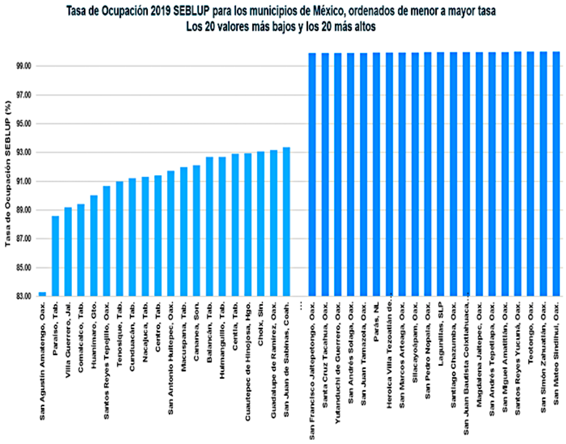


Figure 3: Employment rates obtained from the SAE models by municipality: 20 smallest and 20 largest. Source:INEGI. Cifras laborales para los municipios de México, 2019. Estimación para Áreas Pequeñas. Febrary 2020.

4 Conclusions

The estimation of the economically active population and the employed population show a clear linear relationship with the data obtained in the 2020 Population and Housing Census, which allows us to conclude that the estimates are adequate. However, it is necessary to examine municipalities that have values of 100 percent of economically active population employed, as they could reflect the need to assess the size of the population in order to eliminate the option of such municipalities to have a labour market in which recruitment is done through the search for a job.

Another important element is the characteristic of the economy of the specific state, where higher rates are shown in Oaxaca, Chiapas and Guerrero, which have small municipalities with self-consumption economies.

From the methodological point of view the results can also be improved. As they are, the estimated proportions derived from these models might be inconsistent in the sense that they might not be within the [0, 1] interval, and also the sum of both proportions might exceed one. As indicated by Molina et al (2007), the estimated proportions can be brought to the [0, 1] interval by using logistic models, which relate the logit transformation of the proportions to the auxiliary variables. An additional alternative would be to simultaneously estimate the proportions of unemployed and employed individuals assuming a joint multinomial logit model with random area effects. This model adapts naturally to the characteristics of the problem, solving the inconveniences of previous approaches, and allowing simultaneous model-based estimation of unemployment, employment and inactivity totals, Molina et al. (2007).

References

Dorfman, A. (2018). Towards a Routine External Evaluation Protocol for Small Area Estimation. *International Statistical Review*, 86(2), 259–274

Ibarra-Olivo, E., Acuña, J. and A. Espejo (2021). Estimación de la Informalidad en México a nivel subnacional, Documentos de Proyectos CEPAL. `https://repositorio.cepal.org/bitstream/handle/11362/46789/1/S2000736_es.pdf`.

INEGI (2019). Cifras laborales para los municipios de México, 2018. Estimaciń para Áreas Pequeñas, Agosto, 2019.

INEGI (2019). Encuesta Nacional de Ocupación y Empleo (ENOE). Cómo se hace la ENOE: métodos y procedimientos. 2da ed. 2019.

Molina I., Saei, A. and Lombardia, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society*, series A, 170(4), 975–1000.

Vielma Orozco, E., Vences Rivera, J. and Aguilar Mata, G. (2021). Labor figures for Mexico's municipalities: Small Area Estimation. *Statistical Journal of the IAOS*, 37, 629–640

Flexible Small Area Estimation of Theil Index using Mixture of Beta

Silvia De Nicolò [a*], Maria Rosaria Ferrante [b**] and Silvia Pacei [c**]

[*]Department of Statistical Sciences, University of Padova
[**]Department of Statistical Sciences "P.Fortunati", Alma Mater Studiorum University of Bologna

Abstract

The aim of the paper is to propose a small area estimation model for Theil Index, an entropy-based measure used to quantify economic inequality, industrial concentration and, in general, the disparity related to economic phenomena. We developed an area-level model of its relative index, i.e. Theil index over its maximum, which has a more manageable support between 0 and 1. Classical proposals in area-level context for measures on (0,1) are mostly based on proportions modelling and show limitations when dealing with asymmetric heavy-tailed data, such as in our case. We propose a model with alternative distributional assumptions based on a particular Beta mixture with unconstrained mean modeling, estimated under a Hierarchical Bayes approach. An application to IT-SILC income data is provided, showing that our proposal yields a more flexible framework in comparison with Beta regression with unmatched sampling and linking models.

Keywords— Beta Mixtures, Inequality Mapping, Small Area Estimation, Theil Index.

1 Introduction

In recent year, we are observing an increasing gap in inequality and social exclusion across EU regions. As a consequence, the demand for reliable estimates of economic inequality measures for small areas is growing due to its importance in better planning public and convergence policies. Their estimation in small domains by using income data from household surveys implies that the

[a]silvia.denicolo@phd.unipd.it
[b]maria.ferrante@unibo.it
[c]silvia.pacei@unibo.it

number of units sampled at area level is generally not large enough to obtain reliable estimates. Thus, we have to resort small area estimation techniques, allowing estimators to borrow strength across areas through the use of auxiliary information. See Rao and Molina (2015) for a comprehensive review. The body of literature concerning estimation of inequality measures in small domains is very scarce, comprising Fabrizi and Trivisano (2016) for Gini Index at area level and Tzavidis and Marchetti (2016) for Gini Index and Quintile Share Ratio via M-quantile-based models at unit level. Moreover, inequality can be seen as a multidimensional concept, since different measures are able to capture different aspects of the income distribution. Thus, the estimation of alternative measures may enable a more meaningful and complete overview of the phenomenon.

As opposed to the well known Gini index, Theil index has the advantages to be strongly transfer sensitive, meaning that it react to transfers depending on donor and recipient income levels and it is decomposable among groups. Based on the concept of entropy which applied to income distributions has the meaning of deviations from perfect equality, it pertains to the Generalized Entropy family with parameter $\alpha = 1$:

$$GE(\alpha = 1) = \frac{1}{N} \sum_{k} \frac{x_k}{\mu} \ln \frac{x_k}{\mu}$$

with x_k be a characteristic of interest, for the k-th unit of the finite population, where $x_k \in R^+$, $k = 1, \ldots, N$, and μ its expected value. Since Theil index is defined between 0 and $\log(N)$, we will consider its relative index $RE(1) = GE(1)/\log(N)$ with $GE(1)$ estimated from the survey data with proper weighted estimator and N the true population size. In our estimation strategy, we considered Hierarchical Bayes area level models (Rao and Molina (2015)).

2 Our model proposal

In the context of small area estimation of measures in $(0,1)$, a huge body of literature is dedicated to proportions, implementing Fay-Herriot (Rao and Molina (2015)) and Beta regression models, see Janicki (2020) for a review, with non-linear linking model. The first solution appears restrictive since it may fit values outside the variable support. On the other hand, Beta regression does not provide enough flexibility when facing heavy-tailed, skewed responses and bimodality. Thus our model proposal involves incorporating an alternative distributional assumption on the likelihood of a Hierarchical Bayes area level model, by adopting a beta mixture-based approach. Specifically, we implement the Flexible Beta (FB) distribution proposed by Migliorati et al. (2018), a special mixture of two beta distributions that guarantees a great flexibility and at the same time, great tractability. In fact, the common variance between the two components and their ordered arbitrary means leads it to be identifiable in a strong sense. Let us considered the mean-precision parametrization of the Beta distribution (Ferrari and Cribari-Neto (2004)) such that a generic random variable beta distributed $Y \sim Beta(\mu\phi, (1-\mu)\phi)$, with $E(Y) = \mu$ and

$Var(Y) = \mu(1-\mu)/(\phi+1)$ with $0 < \mu < 1$ and $\phi > 0$ has probability density function $f_B(y;\mu,\phi)$. The FB distribution has pdf

$$f_{FB}(\lambda_1,\lambda_2,\phi,p) = p \cdot f_B(y;\lambda_1,\phi) + (1-p) \cdot f_B(y;\lambda_2,\phi) \tag{1}$$

with $0 < \lambda_2 < \lambda_1 < 1$ distinct ordered means, $0 < p < 1$ and expected value $E(Y) = p\lambda_1 + (1-p)\lambda_2$. Our small area model proposal for y_d, the direct estimator of Theil index and x_d a set of p generic covariates for m small areas is as follows:

$$\begin{cases} y_d|\theta_d \sim FB(\lambda_{1d},\lambda_{2d},\phi_d,p) & \forall d = 1,\ldots,m \\ \text{logit}(\lambda_{2i}) = x_d^T\beta + v_d & v_d \sim N(0,\sigma_v^2) \end{cases} \tag{2}$$

with $\theta_d = E(y_d|\theta_d) = p\lambda_{1d} + (1-p)\lambda_{2d}$ the true parameter value and

$$\phi_d = \frac{\theta_d(1-\theta_d) - Var(y_d|\theta_d)}{Var(y_d|\theta_d) - p(1-p)(\lambda_{1d}-\lambda_{2d})}, \tag{3}$$

where sampling variance $Var(y_d|\theta_d)$ is assumed to be known, as common in literature, in order to allow identifiability. As opposed to the FB regression proposed by Migliorati et al. (2018), the linear predictor does not model directly the mean parameter but rather a mixture component mean, which in this case can be seen as a pure location parameter. This location-modelling approach avoids imposing many vincula on the parameter and simplifies the posterior geometry. In order to carry on the estimation, the parametrization considered is the following: $y_d|\theta_d \sim FB(\tilde{w}_d,\lambda_{2d},\phi_d,p)$ with $\tilde{w}_d = \lambda_{1d} - \lambda_{2d} > 0$. Since estimation requires a variation independent parameter space, we decided to leave λ_{2d},ϕ_d, and p free to assume any value of their support and to constrain $\tilde{w}_d$, whose constrained range is $(0, \min\{1-\lambda_{2d}, \sqrt{Var(y_d|\theta_d)/(p(1-p)}\})$. Thus we model it as $\tilde{w}_d = w \cdot \max\{\tilde{w}_d\}$, with w varying in $(0,1)$ and common to all areas. The separate estimation of sampling variances from data follows a two steps procedure as in Fabrizi et al. (2011). Initially, it is estimated by a proper bootstrap procedure developed taking into account the complex sampling design, using $B = 1000$ repeated samples. Secondly, those estimates are smoothed via a Generalized Variance Function approach in order to reduce bootstrap sampling error. To do so, we derived the variance function of the Theil index as follows.

Proposition 1. *Under the assumption of log-normality of income variable i.e.* $\log(x_{jd}) \sim N(\mu_d,\sigma_d^2)$, *with* $j = 1,\ldots,n_d$ *the individuals and* $d = 1,\ldots,m$ *the areas, the s.r.s. estimator of Relative Theil Index* y_d *has variance function*

$$V(y_d|\theta_d) \cong \frac{2\theta_d^2}{n_d}. \tag{4}$$

Proof. Relative Theil index under log-normal population assumption is

$$\theta_d = \frac{1}{\log(N_d)}\left(\frac{E[x \cdot \log(x)]}{E[x]} - \log(E[x])\right) = \frac{\sigma_d^2}{2\log(N_d)} \tag{5}$$

with N_d population size and since $E[x] = \exp\{\mu_d + \sigma_d^2/2\}$ and $E[x \cdot \log(x)] = (\sigma_d^2 + \mu_d)\exp\{\mu_d + \sigma_d^2/2\}$, with σ_d^2 estimated by $s_d^2 = \sum_{j=1}^{n_d}(\log(x_{jd}) - \hat{\mu}_d)^2/(n_d - 1)$. By applying the normal distribution theory $V(s_d) \cong \sigma_d^2/(2n_d)$ with n_d sample size, and using delta method:

$$V(y_d|\theta_d) = V\left(\frac{s_d^2}{2\log(N_d)}\right) \cong \frac{\sigma_d^4}{2\log^2(N_d)n_d} = \frac{2\theta_d^2}{n_d} \tag{6}$$

where the last right hand side equation is obtained by (5) considering $\sigma_d^2 = 2\theta_d \log(N_d)$. □

Let us assume that $V(y_d|\theta_d) = 2\theta_d^2 \times IF/n_d$ with IF denoting a design-effect variance inflation factor induced by the complex sampling, assumed not to vary across areas, and n_d area sample size under complex sampling. Therefore, considering $\psi = 1/IF$, we introduce the following smoothing model:

$$\frac{2\theta_d^2}{\hat{V}(y_d|\theta_d)_{boot}} = n_d\psi + \varepsilon_d \tag{7}$$

where ε_d are zero-mean and heteroskedastic residuals, estimated via generalized least squares. The smoothed estimator follows from (6) by replacing θ_d with y_d and n_d with $n_d\hat{\psi}$. The following non-informative priors complete the model: $\beta \sim N_p(\bar{0}, \Sigma)$ with Σ diagonal matrix with diagonal $10 \times \bar{1}_p$, $\sigma_v \sim$ Half-Cauchy$(0, 2.5^2)$, $p \sim Unif(0,1)$, $w \sim Unif(0,1)$. We estimated it via Hamiltonian MCMC (`stan`, Carpenter et al. (2017)).

3 Application and Results

An application to assess inequality in Italian NUTS-3 regions through equivalent disposable income data is provided by 2017 EU-SILC survey. Negative income values have been treated by a semi-parametric inverse pareto tail modeling procedure following Finkelstein et al. (2006) and Masseran et al. (2019). As auxiliary variables we considered both fiscal and registry office data related to each of the 107 provinces. In particular: population density, aged dependency ratios, % of foreigners residents, people in higher education ratio, average taxable income, % of residents filling tax forms, % of residents filling tax forms with income lower than/greater than double national median and lastly, Theil index calculated on income classes declared by tax forms.

The Theil index estimator is negatively biased in small samples, direct estimators have been bias-corrected following De Nicolò et al. (2021). We proceeded estimating model in Section (2) and a Beta baseline model having at sampling level $y_d|\theta_d \sim Beta(\theta_d(1-\phi_d), (1-\theta_d)(1-\phi_d))$, and at linking level logit$(\theta_d) \sim N(x_d^T\beta, \sigma_v^2)$. Some diagnostic measures have been used for comparison, as regards goodness-of-fit, `looic`, based on leave-one-out cross-validation Vehtari et al. (2017), whereas a precision improvement measure has been used to evaluate model-based estimators performances, i.e. the Standard Deviation Reduction measure: $SDR(\theta_d) = 1 - [V(\theta_d|\text{data})/\hat{V}(y_d|\theta_d)]^{1/2}$.

As regards Beta model `looic` take value -966.8 (with standard error 16.9), for Flexible Beta model is -992.3 (15.6) showing better goodness of fit. As clear from results set out in Figure 1, our model leads to a greater variability reduction for almost all areas, avoiding negative values, i.e. increases in variability. The FB model provides a standard deviation reductions ranging from 1.6% to 80% with quartiles 32%, 48% and 59%. Moreover, model estimates shows design consistency, i.e. convergence to direct estimators in large samples. Shrinking process of both models is displayed in Figure 2, Beta model estimates present three outliers which have been under-shrank towards the lower tail. This is due to the constrained mean-modelling of the Beta model, highly sensible to covariate values.

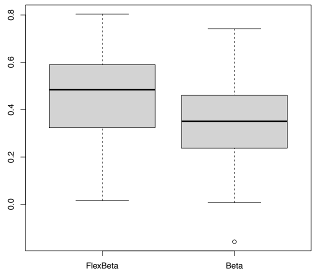


Figure 1: SDR distributions for each area in both models.

4 Concluding Remarks

We proposed a Beta mixture approach for small area estimation of Relative Theil Index, which provides a more flexible framework with respect to Beta regression. Further directions of research involve expanding it to other measures and developing a multivariate context.

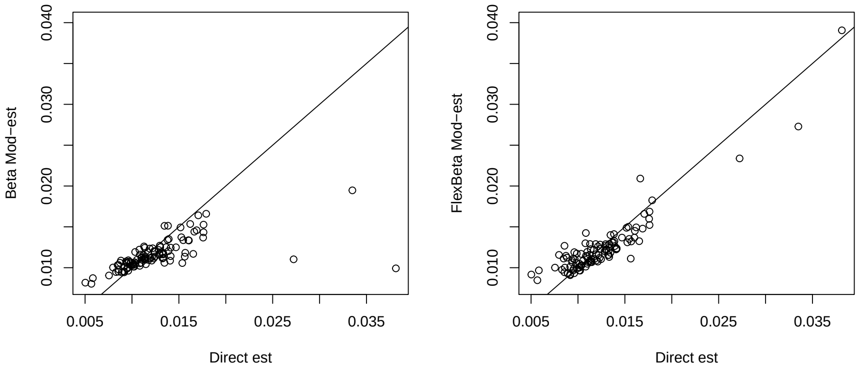


Figure 2: Shrinking process in Beta and Flexible Beta model.

References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: a probabilistic programming language. *Grantee Submission*, 76(1):1–32.

De Nicolò, S., Ferrante, M. R., and Pacei, S. (2021). Mind the Income Gap: Behaviour of Inequality Estimators from Complex Survey Small Samples. *Working Paper.*

Fabrizi, E., Ferrante, M. R., Pacei, S., and Trivisano, C. (2011). Hierarchical Bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. *Computational Statistics and Data Analysis*, 55(4):1736–1747.

Fabrizi, E. and Trivisano, C. (2016). Small area estimation of the Gini concentration coefficient. *Computational Statistics and Data Analysis*, 99:223–234.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815.

Finkelstein, M., Tucker, H. G., and Alan Veeh, J. (2006). Pareto Tail Index Estimation Revisited. *North American Actuarial Journal*, 10(1):1–10.

Janicki, R. (2020). Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. *Communications in Statistics - Theory and Methods*, 49(9):2264–2284.

Masseran, N., Yee, L. H., Safari, M. A. M., and Ibrahim, K. (2019). Power law behavior and tail modeling on low income distribution. *Mathematics and Statistics*, 7(3):70–77.

Migliorati, S., Di Brisco, A. M., Ongaro, A., et al. (2018). A new regression model for bounded responses. *Bayesian Analysis*, 13(3):845–872.

Rao, J. N. and Molina, I. (2015). *Small-area estimation.* Wiley Series in Survey Methodology.

Tzavidis, N. and Marchetti, S. (2016). Robust domain estimation of income-based inequality indicators. *Analysis of Poverty Data by Small Area Estimation*, pages 171–186.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Unit level models on the log-scale: a new Bayesian proposal for poverty mapping

Enrico Fabrizi [a]*, Aldo Gardini [b]** and Carlo Trivisano [c]**

*DISES, Università Cattolica del S. Cuore
**Dipartimento di Scienze Statistiche 'P. Fortunati', Università di Bologna

Abstract

In this paper, we consider the problem of producing estimates of poverty and inequality measures using a Bayesian unit-level small area model, specified on the logarithmic transformation of the equivalised income. In this framework, we extend the classical log-normal model to a finite mixture of log-normal distributions. Moreover, possible negative values are also accomodated. Notoriously, posterior moments for quantities in the original data scale are not necessarily finite under the log-normal model: to solve this problem, we propose a prior specification that guarantees their existence. These methods are applied to Italian data from the EU-SILC survey, complemented with Census information. As domains, we consider sub-population given by administrative provinces by gender.

Keywords— Finite Mixture Model, Generalized Inverse Gaussian, Inequality Indicators, Posterior Moments

1 Introduction

The availability of poverty and inequality indicators estimated for small subsets of large populations enhances the understanding of their distribution across geography and social groups (see Grusky et al., 2006, for a general introduction). In this paper, small area estimation methods (Pratesi, 2016) are exploited to make inference on these quantities, using data from the Italian section of the EU-SILC sample survey complemented by auxiliary information from the Italian population Census. We aim at estimating measures of poverty and inequality obtained as functions of the equivalised income, focusing on the working-age

[a] enrico.fabrizi@unicatt.it
[b] aldo.gardini2@unibo.it
[c] carlo.trivisano@unibo.it

population of Italian administrative provinces classified by gender, since the Italian society is characterized by a marked economic gender divide. Specifically, we target the headcount at-risk-of-poverty rate as a poverty indicator and the quintile share ratio as a measure of inequality.

Concerning the small area literature, we consider unit-level model-based methods and a Bayesian approach to estimation. Since income is typically a positively skewed variable, several authors propose to implement small area methods based on unit-level linear mixed models (and the nested error model, Battese et al. (1988) in particular) specified on the log transformation of the size variable (Elbers et al., 2003; Molina and Rao, 2010, among the others). Unfortunately, the log-normality is not tenable in most applications, and, specifically, it is not in ours. To overcome this problem, we assume a finite mixture of log-normal distributions for the income, an option already discussed in the literature (Lubrano and Ndoye, 2016). Our specification is also related to previous contributions to the small area literature and namely to Chakraborty et al. (2019), who propose a scale mixture of two normal distributions to accommodate outlying residuals in nested error regression models. Notoriously, commonly used priors for the variance components lead to posterior distributions with non-existing moments for several functionals of the un-transformed response variable. To solve this problem, Gardini et al. (2021) propose to use generalized inverse Gaussian distributions (GIG) for variance components. We extended their results to the case of a finite mixture of log-normal distributions, suggesting a prior setting that allows to carry out inference on the target quantities.

2 The proposed model

Before stating the proposed statistical model, some notation needs to be introduced. We target a finite population U of size N, partitioned into D subpopulations $U_1, ..., U_D$ whose sizes $N_1, ..., N_D$ are such that $N = \sum_{d=1}^{D} N_d$. A random sample s of size n is drawn from U, retrieving also the domain-specific sub-samples $s_1, ..., s_D$ with sizes $n_1, ..., n_D$, $n_d \geq 0$, $\sum_{d=1}^{D} n_d = n$. Eventually, the unit-level values of the equivalised income are denoted with y_{di}, $i = 1, \ldots, N_d$, $d = 1, \ldots, D$, whereas $\mathbf{x}_{di} \in \mathbb{R}^p$ contains the observed covariates. The following model is specified:

$$\begin{aligned} &\log(y_{di} + k) = w_{di} = \mathbf{x}_{di}^T \boldsymbol{\beta} + u_d + e_{di},\ d = 1, ..., D;\ i = 1, ..., N_d; \\ &u_d | \tau^2 \sim \mathcal{N}(0, \tau^2),\ e_{di} | \sigma_1^2, \ldots, \sigma_K^2 \sim \sum_{k=0}^{K} \pi_k \mathcal{N}(0, \sigma_k^2); \end{aligned} \tag{1}$$

where K is an integer defining the number of mixed components, $\pi_1, \ldots, \pi_K$ are the weights, with $\sum_k \pi_k = 1$, and $\sigma_1^2, \ldots, \sigma_K^2$ are the variances characterizing the K components. To avoid issues with negative incomes, the constant k must be accordingly fixed. To guarantee the identifiability of the model components, the variances of the components are ordered: $\sigma_1^2 > \cdots > \sigma_K^2$. In the proposed model formulation, the linear predictor slopes as well as the area-specific intercepts are common for all the components, whereas the variances of the individual errors e_{di} are allowed to change from one mixture component to the next. It can be noted that simple log-normal model is a particular case of model (1) when $K = 1$.

The model is specified under the assumption of non-informative sampling, it is fitted on sampled data and, for area d, the observed responses are $y_{di},\ i = 1, \ldots, n_d$; on the other hand, the unsampled ones are $\tilde{y}_{di},\ i = n_d + 1, \ldots, N_d$. To simplify the notation of the subsequent sections, we introduce the vectors $\mathbf{y}_s$ and $\mathbf{w}_s$ containing the responses registered for the sampled units and their (shifted) logarithmic transformation, $\mathbf{X}_s$ is the design matrix with the covariates information.

Monte Carlo Markov Chain (MCMC) methods are used to draw samples from the posterior distributions of the parameters, and indicating with $\tilde{y}_{di}^{(m)}$ the m-th MCMC replicate from the posterior predictive distribution of the unsampled observations, it is possible to define the Bayes predictors as:

$$HCR_d^{(m)}|\mathbf{y}_s = N_d^{-1}\left[\sum_{i=1}^{n_d} \mathbf{1}_{[0;\lambda]}(y_{di}) + \sum_{i=n_d+1}^{N_d} \mathbf{1}_{[0;\lambda]}\left(\tilde{y}_{di}^{(m)}\right)\right]$$

$$QSR_d^{(m)}|\mathbf{y}_s = \frac{\sum_{i=1}^{n_d} y_{di}\mathbf{1}_{[\hat{Q}_{d,0.8}^{(m)};+\infty)}(y_{di}) + \sum_{i=n_d+1}^{N_d} \tilde{y}_{di}^{(m)}\mathbf{1}_{[\hat{Q}_{d,0.8}^{(m)};+\infty)}\left(\tilde{y}_{di}^{(m)}\right)}{\sum_{i=1}^{n_d} y_{di}\mathbf{1}_{[0;\hat{Q}_{d,0.2}^{(m)}]}(y_{di}) + \sum_{i=n_d+1}^{N_d} \tilde{y}_{di}^{(m)}\mathbf{1}_{[0;\hat{Q}_{d,0.2}^{(m)}]}\left(\tilde{y}_{di}^{(m)}\right)},$$

where HCR_d is the headcount ratio, QSR_d the quantile share ration, $\hat{Q}_{d,0.2}^{(m)}$, and $\hat{Q}_{d,0.8}^{(m)}$ represent the first and the fourth quintiles.

A prior distribution for the model parameters must be specified. Starting from the regression coefficients and the mixture weights:

$$\boldsymbol{\beta} \sim \mathcal{N}_p\left(\mathbf{b}_0, \mathbf{V}_0\right), \quad (\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\mathbf{1}); \tag{2}$$

where $\mathbf{1}$ is a K-dimensional vector of ones that allows to specify a uniform distribution on the simplex for the vector of weights.
It can be proved that, under model (1) and priors (2), the moments $\mathbb{E}\left[QSR_d^r|\mathbf{y}_s\right]$ are finite if the priors for the variances associated to unit-level residuals $\sigma_k^2,\ \forall k$ have density function with a term $\exp\left\{-c\sigma_k^2\right\}$ and:

$$c > \frac{r^2}{2}\left[1 + \max_i\left\{\tilde{\mathbf{x}}_{di}^T\left(\mathbf{X}_s^T\mathbf{X}_s + \mathbf{V}_0^{-1}\right)\tilde{\mathbf{x}}_{di}\right\}\right].$$

The prior we choose to fulfill this existence condition is the GIG distribution, a flexible three-parameters distribution with positive support already considered in Gardini et al. (2021) to specify the prior for the variance components a general purpose log-normal mixed model (not allowing for the scale mixture). If $V \sim GIG\left(\lambda, \delta, \gamma\right)$, then its probability density function is:

$$p(v) = \left(\frac{\gamma}{\delta}\right)^{\lambda}\frac{1}{2K_\lambda(\delta\gamma)}v^{\lambda-1}\exp\left\{-\frac{1}{2}(\delta^2 v^{-1} + \gamma^2 v)\right\}\mathbf{1}_{\mathbb{R}^+},$$

where $\lambda \in \mathbb{R}$ is the shape parameter, $\delta \in \mathbb{R}^+$ the scale parameter, and $\gamma \in \mathbb{R}^+$ the tail parameter. The last parameter can be set in order to account for the required existence condition in the prior specification step, keeping

$$\gamma > r\sqrt{\left[1 + \tilde{\mathbf{x}}_{di}^T \left(\mathbf{X}_s^T \mathbf{X}_s + \mathbf{V}_0^{-1}\right) \tilde{\mathbf{x}}_{di}\right]}, \quad \forall d, i.$$

In line with the instructions provided in Gardini et al. (2021), we propose the following independent priors for the variance components:

$$\sigma_k^2 \sim GIG\left(1, 0.01, \gamma_0\right), \; k = 1, \ldots, K; \qquad \tau^2 \sim GIG\left(1, 0.01, \gamma_0\right),$$

with $\gamma_0 = (r+1)\sqrt{\left[1 + \max_{d,i}\left\{\tilde{\mathbf{x}}_{di}^T \left(\mathbf{X}_s^T \mathbf{X}_s + \mathbf{V}_0^{-1}\right) \tilde{\mathbf{x}}_{di}\right\}\right]}$, in order to guarantee the existence the posterior moments of the functionals for any area d. In this way, the priors on the variances are approximately gamma distributions and the induced priors on the intraclass correlation coefficients $\rho_k = \tau^2\left(\tau^2 + \sigma_k^2\right)$ are uniform distributions in the range $(0;1)$.

3 Results from the application

The proposed model is applied to the data from the 2012 Italian sample of the EU-SILC survey, using the equivalised income as response variable and auxiliary information retrieved from the 2011 Italian population census, consisting in counts of the population classified by administrative province, age class (3 levels in the range we consider), sex and education level (4 levels). The combination of administrative province and gender is considered as domain.

Table 1: Goodness-of-fit indicators for the fitted models.

Model	LOOIC (S.E.)	% $CPO < 0.025$	% $CPO < 0.014$
LN	46412 (428)	2.97	2.21
LNM-2	44589 (360)	1.96	1.06
LNM-3	44562 (356)	2.01	1.18

Model (1) is fitted starting from $K = 1$, i.e. the simple log-normal mixed model (LN) and then increasing then the number of components (LNM-K). The performances of the estimated models are compared in Table 1 in terms of LOOIC (Vehtari et al., 2017) and conditional predictive ordinates (CPOs Gelfand, 1996). The mixture model with 2 components improves the LN model, whereas adding a third term does not lead to a considerable change in the goodness-of-fit, and, for this reason, model LNM-2 is chosen for the subsequent considerations.

Figures 1 and 2 shortly illustrate some results. From Figure 1 (left panel), we note that despite HCR is computed on the basis of equivalized income which is the same for all members of a given household, poverty prevalence for women tends to be higher than those for men. This reflects the gender wage gap that

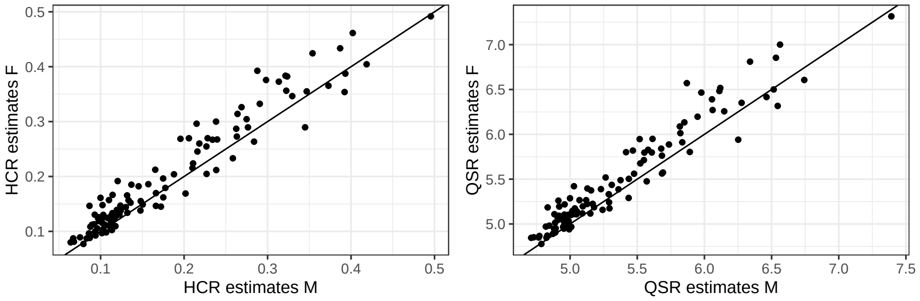


Figure 1: Estimates of the indicators compared by gender.

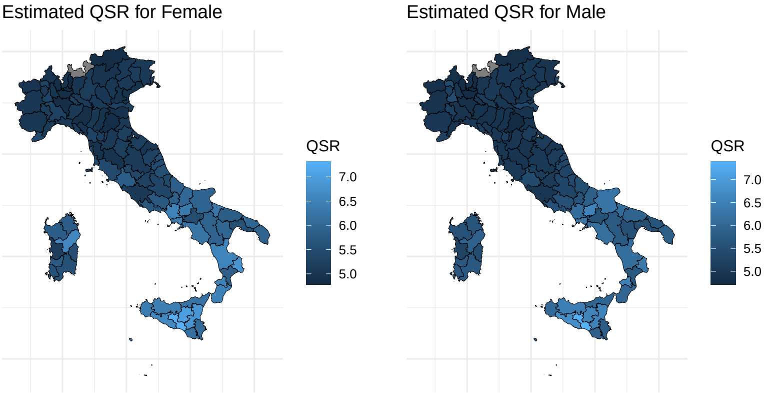


Figure 2: Spatial distribution of QSR.

still characterizes Italy's labour market (Mussida and Picchio, 2014). We note that the wage gender gap translates also in an old-age pension income gap as the latter is positively correlated with wages earned during the working life. Figure 2 highlights the country's North-South divide, with Southern provinces experiencing much larger poverty prevalence. From both figures, we note that poorer areas experiences also larger inequality levels (see also Fabrizi and Trivisano, 2016).

4 Concluding remarks

Beside the application we introduced, we validated our model by means of an exhaustive simulation study, implementing both a model-based and a design-based simulations. The first one highlights that the mixture model does not lose efficiency when the data generating process is actually log-normal, and it outperforms the single components log-normal model when a mixture generates the data. The second study, based on a population generated from a GB2 distribution that mimics the Italian income distribution, confirms the advantages of the mixture model in the estimator performances.

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Chakraborty, A., Datta, G. S., and Mandal, A. (2019). Robust hierarchical bayes small area estimation for the nested error linear regression model. *International Statistical Review*, 87:S158–S176.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1):355–364.

Fabrizi, E. and Trivisano, C. (2016). Small area estimation of the gini concentration coefficient. *Computational Statistics & Data Analysis*, 99:223–234.

Gardini, A., Trivisano, C., and Fabrizi, E. (2021). Bayesian analysis of anova and mixed models on the log-transformed response variable. *Psychometrika*.

Gelfand, A. E. (1996). Model determination using sampling-based methods. *Markov chain Monte Carlo in practice*, pages 145–161.

Grusky, D. B., Kanbur, S. R., and Sen, A. K. (2006). *Poverty and inequality*. Stanford University Press.

Lubrano, M. and Ndoye, A. A. J. (2016). Income inequality decomposition using a finite mixture of log-normal distributions: A bayesian approach. *Computational Statistics & Data Analysis*, 100:830–846.

Molina, I. and Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3):369–385.

Mussida, C. and Picchio, M. (2014). The gender wage gap by education in italy. *The Journal of Economic Inequality*, 12(1):117–147.

Pratesi, M. (2016). *Analysis of poverty data by small area estimation*. John Wiley & Sons.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432.

Defining the sample designs for small area estimation

Piero Demetrio Falorsi [a*], Stefano Falorsi [b*] and Paolo Righi [c*]

[*]Istat, Italy

Abstract

The small area problem is usually considered to be treated via estimation. However, if the domain indicator variables are available for each unit in the population, there are opportunities to be exploited at the survey design stage. This condition is usually met in the business survey context, where the domain indicator variables are available in the business register. The circumstance is respected even in the surveys on households for the geographical domains. Singh, Gambino and Mantel (1994) noted a need to develop an overall strategy that deals with small area problems, involving both planning sample design and estimation aspects. In this framework, it is crucial to control the sample size for each domain of interest so that it is treated as a planned domain at the design stage. It is possible to produce direct estimates with a prefixed level of precision. In general, with a design-based approach to the inference, the presence of sample units in each domain allows one to compute domain estimates, although not always reliably.

In the model-based or model-assisted approach, sample units in each estimation domain allow one to use models with specific small area effects, giving more accurate estimates of the parameters of interest at the small area level (Rao, 2015). Indeed, having sampling units in each domain of interest would also benefit the computation of indirect estimates by enabling a substantial reduction of model bias. Traditional sampling techniques address data disaggregation by oversampling or introducing a more profound stratification. More sophisticated techniques allow improving sampling designs by geographically spreading the sample units (Gräfstorm, Lundström and Schelin, 2012) and diminishing the level of clustering. These approaches would foster reaching segregated or rare subpopulations. In this paper, we consider the problem of estimating the totals Y_d of a variable y for various overlapping domains.

Keywords— Minimum cost solution, balancing equations

[a] International consultant, piero.falorsi@gmail.com
[b] stfalors@istat.it
[c] Parighi@istat.it

1 Introduction

The small area problem is usually considered to be treated via estimation. However, if the domain indicator variables are available for each unit in the population, there are opportunities to be exploited at the survey design stage. This condition is usually met in the business survey context, where the domain indicator variables are available in the business register. The circumstance is respected even in the surveys on households for the geographical domains. Singh, Gambino and Mantel (1994) noted a need to develop an overall strategy that deals with small area problems, involving both planning sample design and estimation aspects. In this framework, it is crucial to control the sample size for each domain of interest so that it is treated as a planned domain at the design stage. It is possible to produce direct estimates with a prefixed level of precision. In general, with a design-based approach to the inference, the presence of sample units in each domain allows one to compute domain estimates, although not always reliably.

In the model-based or model-assisted approach, sample units in each estimation domain allow one to use models with specific small area effects, giving more accurate estimates of the parameters of interest at the small area level (Rao, 2015). Indeed, having sampling units in each domain of interest would also benefit the computation of indirect estimates by enabling a substantial reduction of model bias. Traditional sampling techniques address data disaggregation by oversampling or introducing a more profound stratification. More sophisticated techniques allow improving sampling designs by geographically spreading the sample units (Gräfstorm, Lundström and Schelin, 2012) and diminishing the level of clustering. These approaches would foster reaching segregated or rare subpopulations. In this paper, we consider the problem of estimating the totals Y_d of a variable y for various overlapping domains.

2 Statistical setting

Let U be a target population of size N and let y_i indicate the value of a target variable y of the i-th unit of U. Let $U_d \quad (d = 1, \ldots, D)$ be a particular **subpopulation** of U (being $U_d \in U$) of size N_d and let $\gamma_{id} \quad (i = 1, \ldots, N; \quad d = 1, \ldots, D)$ be the domain membership variable, being $\gamma_{di} = 1$ if $i \in U_d$ and $\gamma_{di} = 0$, otherwise. Let

$$y_{id} = y_i \gamma_{di}.$$

The different domains can overlap, meaning that it is possible $\gamma_{di}\gamma_{d'i} = 1$ for $d = d'$. The latter relationship, implies that the different domains may define alternative partitions of U. For instance, in the business surveys, the domains of interest may be the regions, and the Nace (code of economic activity class). Let

$$Y_d = \sum_{i \in U} y_{id} \quad (d = 1, \ldots, D) \tag{1}$$

be the target domains of interest. With reference to the domain d, we consider a general small-area model

$$y_{id} = \mathbf{x}'_{id}\boldsymbol{\beta} + u_d + e_{id}, \tag{2}$$

where regarding the unit i in the domain d, $\mathbf{x}_{id}$ denotes a vector of auxiliary variables, $\boldsymbol{\beta}$ is a vector of unknown super-population parameters, u_d indicates a random domain effect, and e_{id} a random noise.
The model expectations and variances are $E_M(u_d) = E_M(e_{id}) = 0, \quad V_M(u_d) = \sigma_u^2$, and $V_M(e_{id}) = \sigma^2$.
A sample S of size n is selected without replacement from U with a general sampling design with vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_i, \ldots, \pi_N)'$ of inclusion probabilities. Let

$$\hat{Y}_d = \sum_{i \in S} y_{id} + \sum_{iinU \setminus S} \mathbf{x}'_{id}\hat{\boldsymbol{\beta}} + \hat{u}_d \tag{3}$$

be the small area estimate of Y_d where $\hat{\boldsymbol{\beta}}$ and $\hat{u}_d$ are the sample estimate of $\boldsymbol{\beta}$, and u_d. According to Rao (2015, 7.2.11, pag. 137), we have

$$V_M(\hat{Y}_d) \cong g1(d) + g2(d) + terms \quad of \quad minor \quad order, \tag{4}$$

where

$$g1(d) = N_d^2 \frac{\sigma_u^2 \sigma^2}{n_d \sigma_u^2 + \sigma^2}, \quad \text{and} \tag{5}$$

$$g2(d) = \left[\bar{\mathbf{x}}_{U_d} - \frac{n_d \sigma_u^2}{n_d \sigma_u^2 + \sigma^2} \bar{\mathbf{x}}_{S_d}\right]' \mathbf{A}^{-1} \left[\bar{\mathbf{x}}_{U_d} - \frac{n_d \sigma_u^2}{n_d \sigma_u^2 + \sigma^2} \bar{\mathbf{x}}_{S_d}\right] \text{and} \tag{6}$$

in which n_d is the number of sample units which belong to the small domain U_d and

$$\bar{\mathbf{x}}_{S_d} = \frac{1}{n_d} \sum_{i=1}^{n_d} \mathbf{x}_{id}, \quad \bar{\mathbf{x}}_{U_d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \mathbf{x}_{id},$$
$$\mathbf{A} = \sum_{d \in Par(d)} \sum_{i=1}^{n_d} \mathbf{x}_{id}\mathbf{x}'_{id} \frac{1}{n_d} - \sum_{d \in Par(d)} \frac{n_d \sigma_u^2}{n_d \sigma_u^2 + \sigma^2} n_d \bar{\mathbf{x}}_{S_d} \bar{\mathbf{x}}'_{S_d},$$

$Par(d)$ being the partition of U which includes the domain d. $g1(d)$ is the dominant term of Expression 4 and $g2(d)$ is a term of minor order.

3 Sampling selection

We suppose that the domain indicator variables are available in the sampling frame. We suppose furthermore selecting the sample S by the Cube algorithm (Tillé, 2020) with balancing equations given by

$$\sum_{i\in S}\frac{\mathbf{d}_i}{\pi_1}\cong\sum_{i\in U}\mathbf{d}_i \tag{7}$$

where d_i is a vector of H auxiliary variables for the unit i. Let $\boldsymbol{\gamma}_i = (\gamma_{1i},\ldots,\gamma_{di},\ldots,\gamma_{Di})'$ be the vector of the domain membership indicators, and let $n = (n_1,\ldots,n_d,\ldots n_D)'$ be the vector of the domain sample sizes. If we define the balancing variables in Expression (7) as

$$\mathbf{d}_i = \pi_i\boldsymbol{\gamma}_i, \tag{8}$$

then the sampling selection ensures planned sample sizes, $n_d(d = 1,\ldots,D)$ for each domain. If we define the balancing equations of Expression 7 as

$$\mathbf{d}_i = \left(\pi_i\boldsymbol{\gamma}_i'\frac{1}{n_1}\gamma_{1i}\mathbf{x}_{id}\pi_i,\ldots,\frac{1}{n_D}\gamma_{Di}\mathbf{x}_{iD}\pi_i\right)', \tag{9}$$

then the term $g2(d)$ of Expression 4 vanishes, and the sampling selection ensures the planned sample sizes for each domain. Given the fact that (i) the term $g2(d)$ is of minor order, and (ii) that it vanishes by selecting the sample with appropriate balancing equations, then the main sampling problem is that of defining the minimum cost sample design ensuring that the terms $g1(d)\quad(d = 1,\ldots,D)$ of the model variances $V_M(\hat{Y}_d)$ are lower than pre-fixed thresholds V_d^*. The constrained minimum cost problem may be defined as

$$\begin{cases} Min\left(\sum_{i\in U} C_i\pi_i\right) \\ N_d^2\frac{\sigma^2\sigma_u^2}{\left(\sum_{i\in U}\pi_i\gamma_{di}\right)\sigma_u^2+\sigma^2}\leq V_d^*\,(d=1\ldots,D)\,, \end{cases} \tag{10}$$

where C_i is the unit cost for surveying the unit i. We note that in Problem 10 the variances σ_u^2 and σ^2 are treated as known; in practice they must be estimated. The main issue is to find an algorithmic solution of the problem (10) which represents a non-standard problem. In the context of SSRSWOR sampling, we can define a similar problem, except that the constraints are different. Bethel (1989) invokes the Kuhn-Tucker theorem to show that there exists a solution for the problem of the SSRSWOR sampling. He describes a simple algorithm and discusses its convergence properties. Chromy (1987) develops an algorithm, suitable for automated spreadsheets. The proof of the convergence is given in Falorsi and Righi (2015). Here we describe a modification of the Chromy algorithm to consider the different nature of the constraints in problem (10) with respect to those of the SSRSWOR design. Following Chromy (1987). Let's define the Lagrangian of Problem (10), as:

$$L = \sum_{iinU} C_i \pi_i + \sum_{d=1}^{D} \lambda_d N_d^2 \frac{\sigma^2 \sigma_u^2}{\left(\sum_{i \in U} \pi_i \gamma_{di}\right) \sigma_u^2 + \sigma^2} \quad (i = 1, \ldots, n).$$

Setting $\frac{\partial N}{\partial \pi_i}$, for $i = 1, \ldots, N$, we define a system of N non-linear equation

$$a_i = \boldsymbol{\pi} = 0 \; i = 1, \ldots, N \tag{11}$$

where:

$$a_i(\boldsymbol{\pi}) = \frac{\partial L}{\partial \pi_i} = C_i - \sum_{d=1}^{D} \lambda_d \left[N_d^2 \frac{\sigma^2 \sigma_u^2 \gamma_{di} \sigma_u^2}{\left[\left(\sum_{i \in U} \pi_{i[\alpha]} \gamma_{di}\right) \sigma_u^2 + \sigma^2\right]^2} \right] \quad (i = 1, ldots, n).$$

The values of the Lagrangian λ_d are defined iteratively by the following iteration.

1. **Initialization.** At the first iteration, $\alpha = 0$, we set $\lambda_{d[\alpha]} = 1$.

2. **Calculus.** We find the $\pi_{i[\alpha]}$ $(i = 1, \ldots, N)$ values which solve the following non-linear problem

$$a_i(\boldsymbol{\pi}_{[\alpha_i]}) = C_i - \sum_{d=1}^{D} \lambda_d \left[N_d^2 \frac{\sigma^2 \sigma_u^2 \gamma_{di} \sigma_u^2}{\left[\left(\sum_{i \in U} \pi_{i[\alpha]} \gamma_{di}\right) \sigma_u^2 + \sigma^2\right]^2} \right] \quad (i = 1, \ldots, N).$$

3. **Updating**. We compute

$$g1(d)_{[\alpha]} = N_d^2 \frac{\sigma^2 \sigma_u^2}{\left(\sum_{i \in U} \pi_{[\alpha]} \gamma_{di}\right) \sigma_u^2 + \sigma^2},$$

$$\lambda_{d[\alpha+1]} = \lambda_{d[\alpha]} \left(\frac{g1(d)_{[\alpha]}}{V_d^*} \right)^2 .$$

4. **Iteration.** We set $\alpha = \alpha + 1$. We iterate the steps 2 and 3 till convergence.

References

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology* 15, 47–57.

Chromy, J. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 194–199.

Falorsi, P. D. and Righi, P. (2015). Generalized Framework for Defining the Optimal Inclusion Probabilities of One-Stage Sampling Designs for Multivariate and Multi-domain Surveys. *Survey Methodology*, 41, 215–236.

Grafström A., Lundström N. L. and Schelin L. (2012). Spatially balanced sampling through the pivotal method . *Biometrics*, 68(2), 514–520.

Rao, J. N. K. (2020). *Small Area Estimation.* Wiley, New York.

Singh, M. P., Gambino, J. and Mantel, H. J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3–22.

Tillé, Y. (2020). *Sampling and estimation from finite populations.* John Wiley & Sons, New York.

A Hierarchical Bayesian Approach for Addressing Multiple Objectives in Poverty Research for Small Areas

Stefano Marchetti [a]*, Gaia Bertarelli [b]**, Monica Pratesi [c]* and Partha Lahiri[d]***

*Department of Economics and Management, University of Pisa
**Department of XXX, Sant'Anna School XXX
***Department of Mathematics, University of Maryland

Abstract

Nowadays the information extracted from data should be the key to good policy, therefore, analysts must make the best possible use of all available information. However, data availability often is limited by cost or for other reasons. Consequently, there is the need to use data from different sources. Our goals are to develop hierarchical models and to demonstrate their ability to improve inferences about quantities for which there are meager data. When a hierarchical model can be found to represent the situation properly, analysis of that model often can be used to extract most or all of the relevant information and so provide the best possible estimates. The application considered will include small area estimation in the context of the EU Statistics on Income and Living Conditions. In developing the hierarchical model, we use together survey data and population registers. As for the implementation of the hierarchical model, we propose to use Bayesian methodology assisted by Monte Carlo Markov Chain.

Keywords— poverty mapping, population register, multilevel modelling

1 Introduction

The main goal of this work is to propose a hierarchical model that is able to use data at a different level of aggregation and that come from different sources in

[a] stefano.marchetti@unipi.it
[b] gaia.bertarelli@sssup.it
[c] monica.pratesi@unipi.it
[d] plahiri@umd.edu

order to make inference on the Foster et al. (1984) poverty measures (FGT) at small area level.

In particular we make use of survey data and administrative data, which are collected for many different porpoises, therefore they are available at different level of aggregation. Often, administrative data are aggregated following administrative subdivision of the country, Italy in our application. Italy is divided into 5 repartitions (NUTS 1 level according to the EU nomenclature), 20 regions (NUTS 2), 107 provinces (NUTS 3/LAU 1) and about 8000 municipalities (LAU 2). The goal of our application is the estimation of FGT indexes at the provincial level, in particular poverty incidence. Nevertheless, domains different from administrative boundaries are possible under the proposed framework, e.g. provinces by age class and gender.

The combined use of administrative data and survey data fits the new register-based paradigm of many national statistical offices, where administrative data play a central role in the production of official statistics. This new paradigm requires appropriate models to exploit all the data available in order to produce sound statistics at the small area level.

2 Target paramters and data

Our goal is to obtain FGT indexes at the province level in Italy. Let w_{ijkl} be a wealth variable in region $i = 1, \ldots, R$, province $j = 1, \ldots, D_i$, municipality $k = 1, \ldots, M_{ij}$ and household $l = 1, \ldots, N_{ijk}$ and t be a fixed national threshold that classified poor and non poor households. Then, FGT poverty measure at provincial level is defined as follows:

$$FGT(t)_{ij,\alpha} = \frac{1}{\sum_{k=1}^{M_{ij}} N_{ijk}} \sum_{k=1}^{M_{ij}} \sum_{l=1}^{N_{ijk}} \Big(\frac{t - w_{ijklt}}{t}\Big)^{\alpha} I(w_{ijkl} < t),$$

where $\alpha = \{0, 1, 2\}$ define poverty incidence, intensity and severity respectively. As wealth variable we use the equivalised household income, which is computed as the total available household income divided by the equivalised household size according to the OECD modified scale that assign weight 1 to the first adult, 0.5 to other adults and 0.3 for children (age less than 14).

The equivalised household income is available from the EU Statistics on Income and Living Conditions (SILC) survey, which is conducted yearly by Istat and represent the reference source in the EU for comparative statistics on income distribution and social exclusion. It surveys personal data, income, working status, housing, leisure activities. The 2017 Italian EU-SILC survey has a sample size of about 22 thousand households. It is a two-stage sample design stratified by region and type of municipality. The PSU are the municipalities and the SSU are the households. The survey use a rotating panel each 4 year. More details are available on the Istat website.

Municipalities level data are organised in administrative archives, which are integrated by Istat under the ARCHIMEDE project. The administrative sources used to build ARCHIMEDE microdata are: municipal population registers, tax return registers, central register of pensioners, social security and fiscal sources, social security benefit registers and the population census. Italian population counts about 60 million persons in about 24 million households.

Province level data can be obtained properly aggregating municipality level data, since municipalities are partitions of provinces. However, some data are available only at provincial level, such as the labour force data. This information come the labour force survey (LFS), which is a cross-sectional and longitudinal household sample survey. It provides information about main labour market indicators, broken down by socio-demographic variables. The LFS in Italy follows a rotating sample design where households participate for two consecutive quarters, then they exit for the next two quarters, and finally come back for other two quarters (2 in - 2 out - 2 in rotation). The 2017 LFS in Italy has a sample size of about 250 thousand households, about 600 thousand persons, which guarantee reliable estimates also at the provincial level for what concern annual estimates.

Region specific data can be available from other surveys, but are not considered at this stage in this work.

3 Proposed hierarchical Bayes multilevel model

Hierarchical Bayesian (HB) models have been extensively used in small area estimation, see for example Rao and Molina (2015) for a general review. They can accommodate very complex models based on very simple models as building blocks. Another great advantage of these models is about the estimation of the standard error of the small area HB estimators, which can be found exactly without using approximations. In this framework we can obtain credible intervals and useful summaries from the posterior distributions with practically no additional effort.

In order to obtain reliable estimates of poverty incidence $FGT(t)_{ij,0}$ at provincial level, we propose a three-level cross-sectional model. The three levels are household, municipality and province. Some parameters are defined region specific. The proposed method require to have a transformation $T(w)$ of the wealth variable such that $y = T(w)$ is approximately normal. The mode can be represented as follows:

$$\begin{aligned}
&\text{L.1 } y_{ijkl}|\theta_{ijk}, \boldsymbol{\beta}_i, \sigma_i^2 \sim N(\theta_{ijk} + \mathbf{a}_{ijkl}\boldsymbol{\beta}_i, \sigma_i^2)\\
&\text{L.2 } \theta_{ijk}|\eta_{ij}, \boldsymbol{\gamma}_i, \tau_i^2 \sim N(\eta_{ij} + \mathbf{b}_{ijk}\boldsymbol{\gamma}_i, \tau_i^2)\\
&\text{L.3 } \eta_{ij}|\xi, \boldsymbol{\lambda}, \delta^2 \sim N(\xi + \mathbf{c}_{ij}\boldsymbol{\lambda}, \delta^2),
\end{aligned}$$

where θ_{ijk} is the municipality random effect, $\mathbf{a}_{ijkl}$ are the household level covariates from EU-SILC, η_{ij} is the provincial level random effect, $\mathbf{b}_{ijk}$ are the municipality level covariates from ARCHIMEDE, ξ is a fixed effect and $\mathbf{c}_{ij}$ are the province level covariates from LFS. As a note, c_{ij} covariates are affected by sampling error, which is considered negligible in this work, and then they are treated as true values.

Following Gelman (2015) we use proper informative priors, half-Cauchy for δ, τ_i, σ_i, multivariate normal for $\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i$ and normal for $\xi, \boldsymbol{\lambda}$.

The household level covariate we use is the household size groups: 1 member, 2 members, 3 members, 4 members, 5 or more members. The municipality level covariates are the proportion of persons in age classes (13 to 35, 36 to 65, 66 or more), proportion of male, proportion of persons in 3 type of work contract (dependent, independent, other), median of equivalised taxable income. Note, this last covariate is different from the median equivalised household income estimated from the EU-SILC survey because of a different taxonomy. The province level covariate is the unemployment rate, that is the proportion of persons who don't work while seeking for a job.

An estimate of the unknown quantity $FGT_{ij,\alpha}$ can be obtained as follows:

$$F\bar{G}T(t, \theta_{ijk}, \beta_i, \sigma_i)_{ij,\alpha} = \frac{1}{N_{ij}} \sum_{k=1}^{m_{ij}} \sum_{l=1}^{n_{ijk}} E[g_\alpha(w_{ijkl}) | \theta_{ijk}, \boldsymbol{\beta}_i, \sigma_i^2] \omega_{ijkl},$$

where

$$g_\alpha(w_{ijkl}) = \Big(\frac{t - w_{ijkl}}{t}\Big)^\alpha I(w_{ijkl} < t),$$

m_{ij} are the sampled municipality in province j of region i, ω_{ijkl} is the survey weight for household l in municipality k in province j in region i, n_{ijk} is the sample size in municipality k in province j in region i.

For $\alpha = 0$,

$$\begin{aligned} E[g_\alpha(w_{ijkl}) | \theta_{ijk}, \boldsymbol{\beta}_i, \sigma_i^2] &= \int_{-\frac{\theta_{ijk} + \mathbf{a}_{ijk}\boldsymbol{\beta}_i}{\sigma_i}}^{\frac{\log t - (\theta_{ijk} + \mathbf{a}_{ijk}\boldsymbol{\beta}_i)}{\sigma_i}} \phi(z | \theta_{ijkl}, \sigma_i, \boldsymbol{\beta}_i) dz \\ &= \Phi\Big(\frac{\log t - (\theta_{ijk} + \boldsymbol{a}_{ijk}\boldsymbol{\beta}_i)}{\sigma_i}\Big) - \Phi\Big(-\frac{\theta_{ijk} + \mathbf{a}_{ijk}\boldsymbol{\beta}_i}{\sigma_i}\Big), \end{aligned}$$

where ϕ and Φ are respectively the density function and the distribution function of the standard normal distribution.

The model parameters are estimated using Gibbs sampling by Monte Carlo Markov Chain (MCMC). To obtain stable posterior distribution of model parameters we use a *lasso* penalty. Let H be the number of MCMC samples after burn-in. Let $\theta_{ijk,h}$, $\boldsymbol{\beta}_{i,h}$ and $\sigma_{i,h}$ denote the hth MCMC draw of θ_{ijk}, $\boldsymbol{\beta}_i$ and σ_i, respectively ($h = 1, \ldots, H$). We define the $D_i \times H$ matrix $\mathbf{F}_\alpha$, where the j, h entry is defined as $\mathbf{F}_{(j,h);\alpha} = F\bar{G}T(t, \theta_{ijk,h}, \beta_{i,h}, \sigma_{i,h}))_{ij,\alpha}$.

According to Lahiri and Suntornchost (2018) the matrix $\mathbf{F}_\alpha$ provides samples generated from the posterior distribution of $F\bar{G}T(t, \theta_{ijk}, \beta_i, \sigma_i)_{ij,\alpha}, j = 1 \ldots, D_i, i = 1, \ldots, R$ and so is adequate for solving a variety of inferential problems in a Bayesian way. Lahiri and Suntornchost (2018) suggest three different inferential problems: 1. estimates $\widehat{FGT}(t)_{ij,\alpha}$ of $FGT(t)_{ij,\alpha}$ obtained as the posterior mean of $\mathbf{F}_{(j,h);\alpha}, h = 1, \ldots, H$, and estimates $\widehat{MSE}(\widehat{FGT}(t)_{ij,\alpha})$ of $MSE(\widehat{FGT}(t)_{ij,\alpha})$ obtained as the posterior standard deviation of $\mathbf{F}_{(j,h);\alpha}, h = 1, \ldots, H$; 2. identification of provinces that are out of predefined bounds and 3. identify the worst and best provinces according to FGT indexes. In this work we focus on point 1 only, with $\alpha = 0$. As a remark inference on points 2. and 3. make use of $\mathbf{F}_\alpha$ taking advantage of a unique hierarchical Bayes framework.

4 Application Results

In this section we show the $FGT(t)_{ij,0}$ estimates for 27 provinces in three Italian regions, namely Lombardia, Tuscany and Campania. This choice is due to the availability of ARCHIMEDE data, which have been available to us under an agreement between ISTAT and University of Pisa.

We analyse the model parameters estimates through convergence plot.

We compare the direct estimates with the HB estimates (Figure 1 and Figure 2).

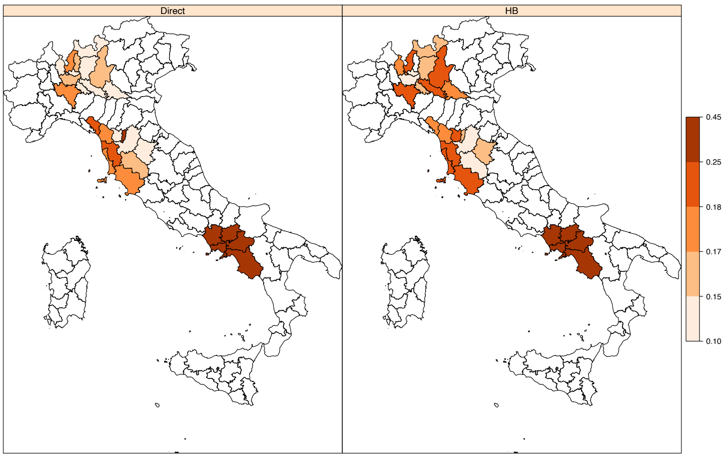


Figure 1: Direct and HB ARPR estimates computed using EU-SILC 2017 at provincial level (NUTS 3)

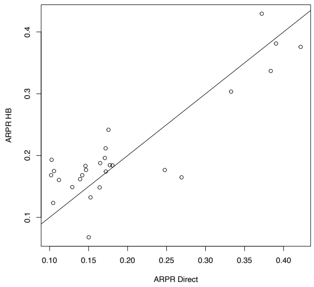


Figure 2: Direct and HB ARPR estimates computed using EU-SILC 2017 at provincial level (NUTS 3).

Model estimates are more reliable than direct ones, with a clear reduction in their coefficients of variation (Figure 3).

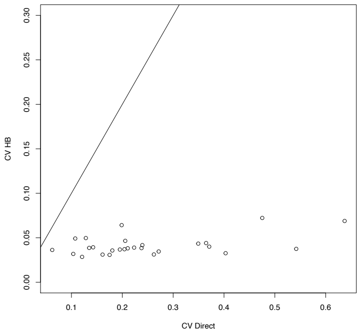


Figure 3: Direct and HB ARPR estimates CVs.

5 Conclusions

In this work we have successfully integrate administrative and survey data at different level of aggregation to obtain posteriors distribution of a multilevel model parameters, which allow different inferential goals. In particular we focus on the incidence of relative poverty, one of the main indicators used by policy makers and stakeholders.

In future works the hierarchical Bayes model can be improved by taking into account the measurement error of auxiliary variables coming from survey data, such as the unemployment rates coming from LFS. Furthermore, the model can be enriched by big data coming for example from google trends, twitter text analysis or supermarket scanner data (which collect price and quantity of retail chains spread across Italy).

References

Foster, J., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, 52:761–766.

Gelman, A. (2015). 3 new priors you can't do without, for coefficients and variance parameters in multilevel regression. Statistical Modeling, Causal Inference, and Social Science blog.

Lahiri, P. and Suntornchost, J. (2018). A general bayesian approach to meet different inferential goals in poverty research for small areas.

Rao, J. and Molina, I. (2015). *Small Area Estimation.* Wiley Series in Survey Methodology. Wiley.

On benchmarking small area estimators when the model is misspecified

Laura Marcis [a*], Maria Chiara Pagliarella [b**] and Renato Salvatore [c*]

[*]University of Cassino and Southern Lazio, Italy
[**]Italian National Institute for Public Policy Analysis

Abstract

One of the main motivations of concern when we apply a small area estimation model is to relate individual area estimates with some direct estimates in a larger area. External and internal benchmarked estimators provide adjusted model-based estimates, in order to agree with that aggregated results. The use of multiple calibration quantities in the benchmarking matrix suggests that the underlying "true" model is misspecified by the actual model equation. We examine the appropriateness of employing the benchmarking matrix to account for omitted variables in the model, through an additional regression term.

Keywords— Fay-Herriot model, benchmarking estimators, model misspecification, augmented model

1 Introduction

In the context of small area estimation, benchmarking is justified by the need for adjusting individual area level estimates to agree with direct estimates of a larger area. The Eblup estimators do not satisfy the benchmarking property, and thus, in the last years, many authors studied a variety of benchmarking techniques, in order to address this issue. In general, these methods rely on some modification of the Eblup by simple adjustments, as for the ratio and the difference benchmarking estimators (Steorts and Ghosh, 2013). Otherwise, an optimal benchmarking estimator that is model unbiased and at the same time satisfies the design-consistency property was obtained by Wang et al. (2008). Bell et al. (2013) give a general result for the optimal estimator in case of

[a] laura.marcis@unicas.it
[b] mc.pagliarella@inapp.org
[c] rsalvatore@unicas.it

multiple benchmarking constraints, by joining together external and internal benchmarking using a common relation. Under model misspecification, Wang et al. (2008) also proposed an augmented model, by inserting a sampling variance model covariate, adjusted by the proportion of units in the corresponding area. Simulation experiments has shown that the augmented model estimator performs well in case of model misspecification, when the omitted variable is correlated with the augmented covariate. Nevertheless, self-benchmarking as in the You and Rao (2002) approach generally ensures efficiency in terms of the MSE, when direct estimates for the larger area suggest a model failure. By a general approach with multiple benchmarking constraints, this paper introduces a benchmarking linear estimator, assuming a model misspecification by an omitted variable factor. The underlying assumption is that direct estimates for the larger area accounts for the true model. Then, we propose an augmented model that incorporates a tentative approach to the model failure. We show that misspecification is proportional to the orthogonal projection of the direct estimate in the subspace of the benchmarking constraints. An application study is reported, in order to introduce the validity of this approach.

2 Theory

Following the Bell et al. (2013) approach to benchmarking small area estimators, as regards the application of the Fay-Herriot model, we say that a) $\theta = X\beta + u$ represents the population area-level parameter model, b) $y = \theta + e$ the sampling model, and, c) $t = W'\theta + \eta$, the benchmarking model by an external random data vector. θ is the $m \times 1$ vector of area parameters, X is the $m \times p$ covariates design matrix, β the $p \times 1$ regression parameters, u is the regression error, y the $m \times 1$ vector of sampling estimates, e is the sampling error, with given $var(e) = R = diag(\psi_1, ..., \psi_m)$. Furthermore, t is the $q \times 1$ vector of benchmarking constraints ($q < m$) to which the area-level estimates must agree, with η the $q \times 1$ related sampling errors. W is a $m \times q$ "benchmarking" matrix, that contains the multiple constraints that links the small area parameters with t. The model variance for θ is $var(y) = Q = \Sigma_u + R$, $\Sigma_u = \sigma_u^2 I_m$. Finally, $cov(u, \eta) = cov(u, e) = 0$, with a chance of a non-zero covariance between sampling errors, i.e. $cov(e, \eta) = C$. Assuming model normality, together with standard Bayesian prior for β, we know that $\Sigma_y = \sigma_\beta^2 XX' + Q$, and, as $\sigma_\beta^2 \longrightarrow \infty$ and by matrix inversion rules, $\Sigma_y^{-1} = Q^{-1}(I - P_X)$. $P_X = X(X'Q^{-1}X)^{-1}X'Q^{-1}$ is the projection matrix onto the subspace by X in the metric of Q^{-1}.

Knowing σ_u^2, and denoting by $\widetilde{\theta}_y = E(\theta|y)$ the best unbiased linear predictor, we have that $\widetilde{\theta}_y = y - RPy$, and $mse(\widetilde{\theta}_y) = var(\widetilde{\theta}_y) = R - RPR$, with P is the projection matrix onto the complement of the column space of X in the metric of Q^{-1}. Assuming the benchmarking model for θ, we get (Bell et al., 2013) $\widetilde{\theta}_{y,t} = E(\theta|y, t)$ as the best linear "adjusted" predictor based on both data sources (y, t):

$$\widetilde{\theta}_{y,t} = E(\theta|y,t) = \widetilde{\theta}_y + cov(\theta,t|y)var(t|y)^{-1}\left\{t - E(t|y)\right\}, \tag{1}$$

$$mse(\widetilde{\theta}_{y,t}) = var(\widetilde{\theta}_y) - cov(\theta,t|y)var(t|y)^{-1}\left\{cov(\theta,t|y)\right\}'. \tag{2}$$

When $var(\eta) = \Sigma_\eta \longrightarrow 0$, $\widetilde{\theta}_{y,t}$ becomes the "externally" benchmarked predictor $\widetilde{\theta}_E = \widetilde{\theta}_y + var(\widetilde{\theta}_y)W[W'var(\widetilde{\theta}_y)W]^{-1}[t - W'\widetilde{\theta}_y]$. Considering $W'\theta = t$ as the projection of the parameter vector θ onto the subspace of W, when we have no external data t and that projection is that relating to y, i.e. $W'y$, the (1) becomes the "internal" benchmarked predictor $\widetilde{\theta}_I = \widetilde{\theta}_y + var(\widetilde{\theta}_y)W[W'var(\widetilde{\theta}_y)W]^{-1}W'(y - \widetilde{\theta}_y)$. In both cases, $\widetilde{\theta}_E$ and $\widetilde{\theta}_I$ verify the benchmarking property $W'\widetilde{\theta}_E = t$ and $W\prime\widetilde{\theta}_I = W'y$, respectively.

In the standard regression theory it is well-known that if for the model in a) we get $E(u|x) \neq 0$, the covariates are said endogeneous in the linear model, i.e. almost one of the explanatory variables is correlated with the regression error u. One of the most important endogeneity problem arises when model misspecification is due to some omitted variables in the equation model. This situation leads in general to the "omitted variable bias" of the fixed-effects estimator, together with an overestimation of the error variance. When important regressors are ignored and the correlation between included and omitted regressors is relevant, the correlation between the covariates and the model error u increases. Conversely, it matters to delete from the model "unimportant" regressors, because they may increase the sampling variance. In large samples, the bias of estimates becomes the major issue (Davidson et al., 2004). Although the standard area level model is mixed linear model, ignoring omitted variables in the regression component of the model, and the consequent unseemly random-area effect variance estimation, may adversely affect the linear predictor. For example, it can be shown that if the true mixed linear model with known model variance Q is $y = X_1\beta_1 + X_2\beta_2 + \pi + Zv + e$, with π as the unobservable omitted vector of fixed effects, and v and e, the random effects and the residual error, respectively, the bias for the fixed-effects estimates of β_1 is $B(\widehat{\beta}_{1,gls}) = A'(I - X_2B')\pi$, $A = Q^{-1}X_1(X_1'Q^{-1}X_1)^{-1}$, $B = PX_2(X_2'PX_2)^{-1}$.

Generally, the presence of omitted variables in the structural linear model significantly correlated with the regression error may be expressed by a model error u, composed of two parts. Denoting by $\theta^m = X\beta + u$ the assumed but incorrect population model, and q the omitted regressor, then $u = q\gamma + v$, where v is the "true" structural regression error. Given the true model θ, we have:

$$\theta = \theta^m + (\theta - \theta^m) = X\beta + q\gamma + v. \tag{3}$$

As q is the unobservable factor in the model for θ, v is uncorrelated with all the covariates $x_1, ..., x_p$, and q. Further, other relations are given, since by the normalization due by the model q has zero mean: $E(u|x) \neq E(u)$, $E(u|q) \neq 0$,$cov(u,q) \neq 0$, $E(v|x,q) = 0$, $E_v(\theta - \theta^m) = q\gamma$. With the true model for θ, and given the benchmarking model c), it is straightforward that $E(t) = W'\theta \neq W'\theta^m$. In the same way, taking the subspace spanned by the columns of W, we observe that $BW'\theta^m \neq BW'\theta = Bt$, $B = W(W'W)^{-1}$, as BW' is the orthogonal projection matrix for W. A proxy-variable solution, say z, for the unobserved factor q, can be assumed as dependent on the difference $(W'\theta - W'\theta^m)$, i.e. $z \propto E(BW'\theta - BW'\theta^m)$. As z becomes a linear regressor for q, we may have in general:

$$q = z\lambda + r = \lambda_0 + \lambda_1 z + r, \tag{4}$$

$E(r|z) = 0, E(r|x,z) = 0$, and, as requested by standard "redundancy" conditions about proxy variables, we can easily check for z that $E(\theta|x,q,z) = E(\theta|x,q)$. Furthermore, given the linear projection $L(q|x,z)$, due to the circumstance that $cov(x,r) = 0$ we may observe that $L(q|x,z) = L(q|z)$. By putting (4) into (3), we get:

$$\begin{aligned} \theta = X\beta + \gamma\lambda_0 1 + \lambda_1 BW'(\theta - X\beta)\gamma + \gamma r + v, \\ (I - \gamma\lambda_1 BW')(\theta - X\beta) = \gamma\lambda_0 1 + \gamma r + v. \end{aligned}$$

Thus:

$$\theta_M = X\overline{\beta} + (I - b_1 BW')^{-1}\epsilon, \tag{5}$$

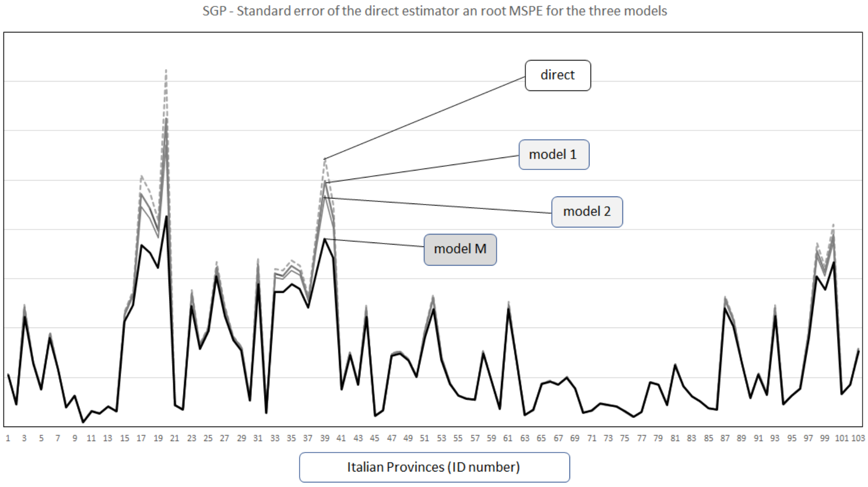
with $\overline{\beta} = \beta_0 + (I - b_1 BW')^{-1} b_0 1$, $\epsilon = \gamma r + v$, $b_0 = \gamma\lambda_0$, $b_1 = \gamma\lambda_1$. By mitigating the omitted factor in the assumed model, equation (5) defines a new model for θ, due to the availability of a proxy variable z. The last by "mirroring" differences in the parameter θ by the projection onto the subspace defined by the benchmarking matrix W. It is straightforward to see for the model (5) that $var(\theta_M) = \sigma^2_\epsilon (I - b_1 BW')^{-1}[(I - b_1 BW')^{-1}]'$, and $cov(\epsilon, \theta_M) = \Sigma_\epsilon (I - b_1 BW')^{-1}$. Although the benchmarking property is always verified for the adjusted predictor (1), whatever the estimate $\widetilde{\theta}_y$ or $\widetilde{\theta}_M$, is certainly interesting to investigate how different models may change estimate of $mse(\theta)$. By the decomposition of the $mse(\widetilde{\theta}) = g_1 + g_2 + g_3$ for the Fay-Herriot model, with the leading term $g_1(\sigma^2_u) = R - RQ^{-1}R = diag(\frac{\sigma^2_u\psi_1}{\sigma^2_u+\psi_1}, ..., \frac{\sigma^2_u\psi_m}{\sigma^2_u+\psi_m})$, it is straightforward to note that $g_{1,i}(\sigma^2_v) < g_{1,i}(\sigma^2_u)$, $\forall i$, $i = 1, ..., m$, when $\sigma^2_v < \sigma^2_u$. Further, with $Q = diag(\sigma^2_v + \psi_1, ..., \sigma^2_v + \psi_m)$, $Q_m = diag(\sigma^2_u + \psi_1, ..., \sigma^2_u + \psi_m)$, and following standard matrix inversion rules, $Q_m^{-1} = Q^{-1} - S^{-1}$, $S = diag[\frac{(\sigma^2_v+\psi_1)(\sigma^2_u+\psi_1)}{\sigma^2_u-\sigma^2_v}, ..., \frac{(\sigma^2_v+\psi_m)(\sigma^2_u+\psi_m)}{\sigma^2_u-\sigma^2_v})]$, it can be shown that:

$$\begin{aligned} tr[mse(\widetilde{\theta}_M)] &= tr\left\{var[\widetilde{\theta}_y(\sigma^2_v)] + R(Q_m^{-1} + S^{-1})P_{(I-P_X)q}R\right\} \\ &< tr\left\{var[\widetilde{\theta}_y(\sigma^2_u)] + R(Q_m^{-1} + S^{-1})P_{(I-P_X)q}R\right\} = tr[mse(\widetilde{\theta}_y)]. \end{aligned}$$

Here the projection matrix of the model (3) is partitioned by the decomposition $P_{X,q} = P_X + P_{(I-P_X)q}$, where the matrix $P_{(I-P_X)q}$ projects vectors onto the space spanned by the columns of $(q - P_X q)$.

3 Estimation, application study, conclusions

The model in (3), equipped with the proxy-variable solution expressed by (5), contains the parameters γ, λ_0, and λ_1. While the last two are regression parameters, the γ parameter gives the level of dependence of the "true" model respect to the omitted factor. Our proposal is to firstly give an estimate of γ, in order to provide more easily the estimates of the remaining two parameters, λ_0 and λ_1. Starting from the error in the assumed model for θ (i.e. θ^m), $u = q\gamma + v$, first give the Fay-Heriott model estimate $\widetilde{u}$. Remembering that in our belief the random effects in the incorrect model contain an omitted factor q, we fit a standard regression model d) $\widetilde{u} = \gamma(y - \widetilde{y}) + \epsilon$ to the predicted data, to thus obtain an estimate $\widehat{\gamma}$. The underlying condition to verify is the following: if we have no omitted factors in the model, the contribution to the random effect of $(y - \widetilde{y})$ is random with zero mean, i.e. $u \equiv v$ in the population model. With the estimate $\widehat{\gamma}$ of γ, standard estimation methods are available, like REML or ML in case of normal distribution. If we set in the model (5) $\lambda_0 = b_0 = 0$ as a no-intercept model for q in (4), it is possible to provide the estimation of the model (5), quite similar to a simultaneous autoregressive spatial Fay-Herriot model. We have indeed $\rho_t \equiv b_1 = \widehat{\gamma}\lambda_1$, and $D \equiv BW'$, $b_1 \neq 1$, being ρ_t the time-autocorrelation parameter to estimate, and D a standardized proximity matrix that define both a spatial model, taken as an example of comparison. With $b_0 \neq 0$ and fixing $\widehat{\gamma}$, we propose to alternate by the chosen estimation method both b_1 and b_0, in turn. In the data study we present an application of the Fay-Herriot area-level model to official Agricultural Census Data (Survey structure and production of Italian farms (year 2007). The dataset contains 103 observations (the small areas in our context) that are the Italian provinces. The target variable is the mean of standard gross profit (SGP). The selected



auxiliary variables are the mean of irrigable area (IRR) and the mean of the number of worked days (DAYS). We use restricted maximum likelihood estimators (REML) at the several step of estimation. To illustrate the experiment, it is important to know that the SGP is a composite variable, depending on several economic features of farms. It depends on IRR and DAYS, but also on other farm characteristics. The "larger areas" are the 20 administrative Region (NUTS2 level), which includes several Provinces (NUTS3 level). The application consists on comparing estimates and their mean squared prediction errors (MSPE) only, respect to the standard Fay-Herriot model. We use a benchmarking matrix for a simple "internal" benchmarhing design, $W_{103\times 1}$, that reports some weights to agree small area estimates with regional available direct estimates of the mean of SGP of the farms. The plot illustrates results from three models: "model 1", the model that has only the DAYS covariate, "model 2" with both "DAYS" and "IRR" covariates, and "model M", the presented model in (5). An estimate of γ by the model in d), $\widehat{\gamma} = 3.969$, together with a no-intercept proxy-variable model ($\lambda_0 = b_0 = 0$) is used, that gives the REML estimate $b_1 = \widehat{\gamma}\lambda_1 = 0.898 \longrightarrow \widehat{\lambda}_1 = 0.226$.

The plot reports differences in the predicted root mean squared errors for the three model investigated. It is interesting to note that the progressive inclusion of covariates (from model 1 to model 2), tends to what it is the possible "true" model, that build the SGP dependent variable with a lot of variables, included DAYS and IRR. The application presented is only a first step respect to what may be the potential of the introduced model (5). A possible definitive proof about the quality of the proposed approach to the benchmarking issues in small area estimation, will probably be represented by simulation experiments. In our opinion, one of the most important features of the approach is given by exploiting the benchmarking matrix as appears in the augmented factor in the model. This matrix viewed as an "implicit bringer" of information, toward the identification of a "true" model for the small area statistics.

References

Bell, W. R., Datta, G. S., and Ghosh, M. (2013). Benchmarking small area estimators. *Biometrika*, 100(1):189–202.

Davidson, R., MacKinnon, J. G., et al. (2004). *Econometric theory and methods*, volume 5. Oxford University Press New York.

Steorts, R. C. and Ghosh, M. (2013). On estimation of mean squared errors of benchmarked empirical bayes estimators. *Statistica Sinica*, pages 749–767.

Wang, J., Fuller, W. A., and Qu, Y. (2008). Small area estimation under a restriction. *Survey methodology*, 34(1):29.

You, Y. and Rao, J. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30(3):431–439.

Empirical Best Prediction for Small Area Estimation of categorical variables using Finite Mixtures of Multinomial Logistic Models

M.G. Ranalli [a*], M.F. Marino [b**], N. Salvati [c***] and M. Alfò [d****]

[*]Dept. of Political Science, Università degli Studi di Perugia, Italy
[**]Dept. of Statistics, Computer Science, Applications, Università degli Studi di Firenze, Italy
[***]Dept. of Economics and Management, Università di Pisa, Italy
[****]Dept. of Statistical Science, Sapienza Università di Roma, Italy

Abstract

Many survey variables are categorical in nature and SAE methods based on generalised linear mixed models represent a frequent tool of analysis for prediction. Jiang (2003) developed an Empirical Best Prediction (EBP) method for responses in the Exponential Family, based on the use of area-specific, Gaussian, random effects. However, a major drawback of this approach is the computational burden required to derive estimates, compute the EBP and, in particular, provide the corresponding measure of reliability. Here, we introduce a semiparametric EBP for categorical outcomes by extending the approach proposed by Marino et al. (2019) for univariate responses belonging to the Exponential Family of distributions. This approach leaves the mixing distribution (that is, the distribution of the area-specific random effects) unspecified and estimate it from the observed data via a NonParametric Maximum Likelihood approach. This estimate is known to be a discrete distribution defined over a finite number of locations and leads to the definition of a finite mixture specification. Finite sample properties of the proposal are tested via a simulation study.

1 Introduction

Quite often survey variables are categorical and when researcher's interest entails prediction, SAE methods based on the use of generalised linear models identify a rather

[a]giovanna.ranalli@unipg.it
[b]mariafrancesca.marino@unifi.it
[c]nicola.salvati@unipi.it
[d]marco.alfo@uniroma1.it

standard tool of analysis. For responses belonging to the univariate Exponential Family of distributions, Jiang (2003) developed an Empirical Best Prediction (EBP) method. More recently, Boubeta et al. (2016, 2017) derived the EBP and the corresponding (second-order) approximation to the MSE under an area-level mixed Poisson model for small area counts, while Hobza and Morales (2016) specifically focused on the development of an EBP for small area proportions under the unit-level mixed logistic model considered in Jiang (2003). An extension of this latter approach to deal with longitudinal responses was recently proposed by Hobza et al. (2018).

These proposals are based on unit-level mixed models and area-specific random effects which are assumed to be i.i.d. according to a Gaussian density. One of the drawbacks associated with this parametric approach entails the computational cost of deriving parameter estimates, compute the EBP and, in particular, provide the corresponding measure of reliability. This is due to the need of approximating (possibly) multiple integrals without a closed form expression. Monte Carlo integration and parametric bootstrap are frequently considered to obtain an approximation. To avoid computational issues, ad hoc alternatives, mainly based on plug-in predictors, were proposed in Molina et al. (2007); Saei and Taylor (2012). In particular, Molina et al. (2007) suggest the use of a SAE model in which the area-specific random coefficients are assumed to be Gaussian and shared by all model equations; Saei and Taylor (2012) relaxed this latter assumption by considering area- and category-specific random coefficients to enhance model flexibility. In both cases however, plug-in predictions are considered to estimate small area proportions for response' categories.

In this paper, we extend the semiparametric best predictor approach introduced by Marino et al. (2019) for univariate responses in the Exponential Family to the multinomial case. We leave the distribution of the area-specific random effects unspecified and estimate it from the observed data via a NonParametric Maximum Likelihood approach (NPML Laird, 1978). This estimate is known to be a discrete distribution defined over a finite number of locations leading to a finite mixture model. Such an approach offers a number of advantages. First, it allows us to obtain an EBP and not a plug-in approximation and avoid unverifiable assumptions on the random effect distribution; second, since mixture parameters are directly estimated from the data and are completely free to vary over the corresponding support, extreme and/or asymmetric departures from the homogeneous model can be easily accommodated. Last and more important, the discrete nature of the mixing distribution allows us to avoid integral approximations and considerably reduces the computational effort.

The paper is organised as follows. In Section 2 we illustrate the method, while in Section 3 we report the results of a simulation study that explores its performance. Finally, Section 4 provides concluding remarks and hints at ongoing work on MSE estimation and application to real data on employment indicators.

2 Semiparametric EBP for categorical data

Let $Y_{ij} = (Y_{ij1}, \dots, Y_{ijK})'$ denote a multinomial response for unit j belonging to small area i ($i = 1, \dots, m, j = 1, \dots, N_i$), whose generic element Y_{ijk} is equal to 1 if the ij-th unit is in the k-th category ($k = 1, \dots, K$), and is equal to 0 otherwise; furthermore, we have that $\sum_{k=1}^{K} Y_{ijk} = 1$. Let $\alpha_i = (\alpha_{i1}, \dots, \alpha_{iK})'$ be an area-specific random vector associated to area i, x_{ij} denote a p-dimensional vector of covariates, with X_i the matrix of covariates for the i-th small area.

We assume that, conditional on α_i, responses for units in the i-th small area $Y_i = \{Y_{i1}, \dots, Y_{iN_i}\}$ are independent each other and each element in Y_{ij}, say Y_{ijk}, is influenced by the corresponding element in α_i only, that is α_{ik}. In particular, we assume that, conditional on α_i, responses Y_{ij} follow a multinomial distribution, with parameters 1 and probability elements $p_{ij} = (p_{ij1}, \dots, p_{ijK})$; these latter are modeled according to the following multinomial logit specification:

$$\theta_{ijk} = \log \frac{p_{ijk}}{p_{ij1}} = \alpha_{ik} + x_{ij}'\beta_k, \quad k = 2, \dots, K. \tag{1}$$

Here, β_k denotes a p-dimensional vector of fixed model parameters that describes the effect of observed covariates on the multinomial logit transform of p_{ijk}.

Our aim is that of predicting small area proportions $\bar{Y}_i = (\bar{Y}_{i1}, \dots, \bar{Y}_{iK})'$, with

$$\bar{Y}_{ik} = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ijk},$$

using model (1) and assuming that responses Y_{ij} are observed for sampled units only $(i = 1, \dots, m, j \in s_i)$, while covariates x_{ij} are available at the population level $(i = 1, \dots, m, j = 1, \dots, N_i)$. To this aim, we extend the semiparametric best prediction approach introduced by Marino et al. (2019) for univariate outcomes to the multinomial framework. In detail, we define the semiparametric best predictor for the quantity $p_i = (p_{i1}, \dots, p_{ik})'$, with

$$p_{ik} = \frac{1}{N_i} \sum_{j=1}^{N_i} p_{ijk} \tag{2}$$

by assuming that the random coefficients α_i follow a discrete distribution defined over the finite set of locations $\xi_g = (\xi_{g1}, \dots, \xi_{gK})'$ with masses

$$\pi_g = \Pr(\alpha_i = \xi_g) = \Pr(\alpha_{i1} = \xi_{g1}, \dots, \alpha_{iK} = \xi_{gK}),$$

where $\pi_g > 0$, $\sum_{g=1}^{G} \pi_g = 1$. This approach is similar to that detailed by Alfò et al. (2021) in the context of multivariate longitudinal measures and leads to the definition of the following model likelihood:

$$L(\cdot) = \sum_{i=1}^{m} \sum_{g=1}^{G} \left[\prod_{j \in s_i} f(y_{ij} \mid \alpha_i = \xi_g, X_i) \right] \pi_g, \tag{3}$$

where $f(y_{ij} \mid \alpha_i = \xi_g, X_i)$ denotes the density for the observed responses of the $j-th$ unit belonging to the i-th small area, conditional on $\alpha_i = \xi_g$. That is, it corresponds to the Exponential Family density with canonical paramerer

$$\theta_{ijkg} = \log \frac{p_{ijkg}}{p_{ij1g}} = \xi_{gk} + x_{ij}'\beta_k.$$

Turning back to the problem of estimating p_i, we have that the semiparametric best predictor of each component p_{ik} in p_i is given by

$$\begin{aligned}\tilde{p}_{ik}^{\text{sp-BP}} &= \sum_{g=1}^{G} p_{i\cdot kg} \frac{\exp\left[\sum_{j\in s_i}\sum_{k=1}^{K-1} y_{ijk}\theta_{ijkg} - \sum_{j\in s_i}\log\left(1+\sum_{k=1}^{K-1} e^{\theta_{ijkg}}\right)\right]\pi_g}{\sum_{l=1}^{G}\exp\left[\sum_{j\in s_i}\sum_{k=1}^{K-1} y_{ijk}\theta_{ijkl} - \sum_{j\in s_i}\log\left(1+\sum_{k=1}^{K-1} e^{\theta_{ijkl}}\right)\right]\pi_l} \\ &= \sum_{g=1}^{G} p_{i\cdot kg} \frac{\exp\left[\sum_{k=1}^{K-1}\alpha_{gk}y_{i\cdot k} - \sum_{j\in s_i}\log\left(1+\sum_{k=1}^{K-1} e^{\theta_{ijkg}}\right)\right]\pi_g}{\sum_{l=1}^{G}\exp\left[\sum_{k=1}^{K-1}\alpha_{lg}y_{i\cdot k} - \sum_{j\in s_i}\log\left(1+\sum_{k=1}^{K-1} e^{\theta_{ijkl}})\right)\right]\pi_l}\end{aligned}$$

where $y_{i\cdot k} = \sum_{j\in s_i} y_{ijk}$ and $p_{i\cdot kg} = N_i^{-1}\sum_{j=1}^{N_i} p_{ijkg}$. By letting

$$\tau_{ig(y_{i\cdot})} = \frac{\exp\left[\sum_{k=1}^{K-1}\alpha_{gk}y_{i\cdot k} - \sum_{j\in s_i}\log\left(1+\sum_{k=1}^{K-1} e^{\theta_{ijkg}}\right)\right]\pi_g}{\sum_{l=1}^{G}\exp\left[\sum_{k=1}^{K-1}\alpha_{lk}y_{i\cdot k} - \sum_{j\in s_i}\log\left(1+\sum_{k=1}^{K-1} e^{\theta_{ijkl}})\right)\right]\pi_l},$$

the sp-BP of p_{ik} is given by $\tilde{p}_{ik}^{\text{sp-BP}} = \sum_{g=1}^{G} p_{i\cdot kg}\,\tau_{ig(y_{i\cdot})}$. In matrix form, we may re-write the above problem as

$$\tilde{p}_i^{\text{sb-BP}} = P'_{i[1:G]}\tau_{i(y_{i\cdot})} \tag{4}$$

where

$$\tilde{p}_i^{\text{sb-BP}} = \begin{pmatrix} p_{i1}^{\text{sb-BP}} \\ \vdots \\ p_{iK}^{\text{sb-BP}} \end{pmatrix}, \quad P_{i[1:G]} = \begin{pmatrix} p_{i\cdot 11} & \cdots & p_{i\cdot K1} \\ p_{i\cdot 12} & \cdots & p_{i\cdot K2} \\ \vdots & \vdots & \vdots \\ p_{i\cdot 1G} & \cdots, & p_{i\cdot kg} \end{pmatrix}, \quad \tau_{i(y_{i\cdot})} = \begin{pmatrix} \tau_{i1(y_{i\cdot})} \\ \vdots \\ \tau_{iG(y_{i\cdot})} \end{pmatrix}.$$

The corresponding sp-EBP, denoted by $\hat{p}_i^{\text{sb-EBP}}$, is obtained by plugging ML estimates of model parameters into expression (4):

$$\hat{p}_i^{\text{sb-EBP}} = \hat{P}'_{i[1:G]}\hat{\tau}_{i(y_{i\cdot})}. \tag{5}$$

ML estimates of model parameters can be obtained using the EM algorithm. Note that, for a given choice of G, the prediction problem is solved by rewriting the general integral used in the parametric approach as a sum over G components. This leads to a substantial save in computational complexity as often integral approximation by parametric methods leads to a sum over a much larger number of locations.

3 Simulation study

To evaluate the empirical properties of the proposal, we conducted a model-based simulation study considering $T = 1,000$ samples. Multinomial population data are generated considering $m = 100$, 200, 500 small areas under the proposed modelling assumptions. From such a population, sample data are selected via a simple random sampling without replacement within each area. The population and the sample sizes are assumed to be constant across areas and are fixed to $N_i = 100$ and $n_i = 10$, respectively. For each unit j in small area i, we generate the target variable $Y_{ij}, i = 1,\ldots,m, j =$

$1,\dots,N_i$, from a Multinomial distribution with parameters 1 and p_{ij}, with components of p_{ijk} defined as

$$p_{ijk} = \frac{e^{\alpha_{ik}+x'_{ij}\beta_k}}{1+\sum_{k=2}^{K} e^{\alpha_{ik}+x'_{ij}\beta_k}}, \quad k = 2,3,$$

with $\beta_2 = 0.5$, $\beta_3 = -0.5$, $x_{ij} \sim \text{Unif}(-1, b_i)$, and $b_i = i/8$, $i/16$, $i/48$ for $m = 100, 200$ and 500, respectively. The simulation settings are those used in González-Manteiga et al. (2007) and suitably adapted to the multinomial case. As regards the random coefficients $\alpha_{ik}, k = 2,3$, these are generated from both a multivariate Gaussian and a mixture of multivariate Gaussian density, with both uncorrelated ($\rho_{\alpha_2\alpha_3} = 0$) and correlated components ($\rho_{\alpha_2\alpha_3} \neq 0$).

Starting from parameter estimates computed via a NPML approach, we derived the sp-EBP for p_{ik} and compared results with those obtained by considering the approach detailed by Molina et al. (2007) and Saei and Taylor (2012) in terms of absolute bias and root mean square error over areas (AB_i and $RMSE$).

Simulation results (not reported here for reasons of space) highlight that, in the presence of multivariate Gaussian random coefficients (besides their correlation) the proposal performs better than that by Molina et al. (2007) and slightly worse than that by Saei and Taylor (2012), both in terms of bias and RMSE. On the other side, when random coefficients are generated from a mixture of multivariate Gaussian densities, the semiparametric EBP returns predictions with a much lower bias and RMSE than competitors.

4 Conclusions

In this work, we extended the semiparametric EBP introduced by Marino et al. (2019) for general responses in the univariate Exponential Family to the multinomial framework. Multivariate, area-specific random coefficients, with category-specific components, are considered to account for dependence between outcomes from the same small area. The corresponding multivariate distribution is left unspecified and estimated through the observed data via a NPML approach, leading to the definition of a finite mixture model likelihood.

A large scale simulation study was performed to assess the quality of predictions obtained thanks to the proposed approach. Here, the chosen scenarios represent two extreme situations; we expect that, in real applications, the random coefficient distribution lies in between them. Results of such a study highlight good performance of the proposal so that is seems a promising approach to consider when dealing with the problem of predicting small area proportions for categorical outcomes.

An analytic expression for the MSE of the sp-EBP is available and we are currently working on its implementation. In particular, the quality of predictions obtained via $\hat{p}_i^{\text{sp-EBP}}$ can be evaluated through the following analytic MSE expression:

$$\text{MSE}(\hat{p}_i^{\text{sp-EBP}}) = E_\alpha[p_i p_i'] - E_y[\tilde{p}_i^{\text{sp-BP}}(\tilde{p}_i^{\text{sp-BP}})'] + E_y[(\hat{p}_i^{\text{sp-EBP}} - \tilde{p}_i^{\text{sp-BP}})(\hat{p}_i^{\text{sp-EBP}} - \tilde{p}_i^{\text{sp-BP}})'].$$

As an alternative, a bootstrap approach may be adopted.

Finally, a main goal of the project is the application of the proposal to obtain estimates of employment, unemployment, and inactive counts and proportions using data from the Italian Labor Force Survey for Local Labor Market Areas (LLMAs). LLMAs are 611 unplanned domains obtained as clusters of municipalities and defined at the Census on the basis of daily working commuting flows. In this context, direct survey estimates cannot be computed and/or published for most of the LLMAs due to the presence of many out-of-sample areas and small sample sizes. In this context, indirect, model-based, small area estimators are adopted by ISTAT to produce official yearly estimates of labor market indicators for the Italian LLMAs, separately for employed and unemployed. A multivariate perspective would certainly provide more insights and grant internal coherence of the final estimates.

References

Alfò, M., Marino, M. F., Ranalli, M. G., Salvati, N., and Tzavidis, N. (2021). M-quantile regression for multivariate longitudinal data with an application to the millennium cohort study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(1):122–146.

Boubeta, M., Lombardía, M. J., and Morales, D. (2016). Empirical best prediction under area-level Poisson mixed models. *Test*, 25(3):548–569.

Boubeta, M., Lombardía, M. J., and Morales, D. (2017). Poisson mixed models for studying the poverty in small areas. *Computational Statistics & Data Analysis*, 107:32–47.

González-Manteiga, W., Lombardía, M., Molina, I., Morales, D., and Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics & Data Analysis*, 51:2720–2733.

Hobza, T. and Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. *Journal of Official Statistics*, 32:661–692.

Hobza, T., Morales, D., and Santamaría, L. (2018). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *TEST*, 27:270–294.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111:117–127.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73:805–811.

Marino, M. F., Ranalli, M. G., Salvati, N., Alfò, M., et al. (2019). Semiparametric empirical best prediction for small area estimation of unemployment indicators. *The Annals of Applied Statistics*, 13(2):1166–1197.

Molina, I., Saei, A., and Lombardía, M. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society: Series A*, 70:265–283.

Saei, A. and Taylor, A. (2012). Labour force status estimates under a bivariate random components model. *Journal of the Indian Society of Agricultural Statistics*, 66(1):187–201.

Design-based small area estimation: an application to the DHS surveys

Ruilin Ren

ICF International, Rockville, MD,USA

Abstract This study explores design-based small area estimation methods using Demographic and Health Survey (DHS) data collected by The DHS Program, an international project funded by United States Agency for International Development (USAID). The DHS surveys are household based two-stage cluster surveys that provide key survey indicators for a country's first level administrative unit, or region. The DHS Program faces increasing requests from host countries for sub-regional indicator estimates for policymaking and development planning purposes. Increasing sample size is not usually feasible for meeting this need. One solution is to use small area estimation techniques to produce reliable estimation for sub-regions. This study explores a method that "borrows" strength from within the target survey or from similar surveys conducted recently in the same country. The idea is to create a survey "domain" covering the small area by pooling clusters close to the small area within the target survey or from similar surveys conducted in recent years. "Close" means geographically, or in time and space, or using other measures such as demographic, social, religious, cultural, or economic measures. Using design-based domain analysis tools, the calculation of parameter estimation, variance estimation and confidence intervals for small areas is straightforward. This study uses data from the 2010 and 2014 Rwanda DHS surveys and the proposed methods to produce district level total fertility rates (TFR) which were not provided in the survey reports due to insufficient sample sizes at the district level.

Key words: small area estimation; design-based; survey domain; total fertility rate (TFR)

ruilin.ren@icf.com

1 Introduction

The Demographic and Health Surveys (DHS) Program is an international project funded by the United States Agency for International Development (USAID). The DHS Program has collected, analysed, and disseminated high quality data on population, health, HIV, malaria, and nutrition and other topics through more than 400 national surveys in over 90 countries since 1984. The DHS Program receives more and more requests from host countries to produce estimates of key DHS indicators at the sub-regional/district level. Direct estimates, especially for total fertility rates (TFR) and childhood mortality rates (CMR) which need a very large sample size, are not feasible because of the cost and possible adverse effect on data quality, especially when there is a large number of sub-regions. One solution is to use small area estimation techniques to produce reliable sub-regional estimates. This research explores a design-based methodology, by “borrowing strength” from data collected in a single target survey or in similar surveys conducted in recent years. The idea is to create a survey domain which is the nearest neighbour to the small area by pooling clusters close to the small area geographically, in time and space, or by using other measures such as demographic, social, religious, cultural, and economic measures within one survey or from similar surveys conducted in recent years. A survey domain created in this way is easy to analyse using design-based domain analysis tools to calculate parameter estimates, variance estimates and confidence intervals. Data from DHS surveys are typically more reliable and often more timely than data from other sources such as censuses or administrative records. A calibration procedure can be applied to small area estimates to ensure that they can be aggregated to the regional level and match the regional level estimates of the target survey. This consistence property is desirable for small area estimation.

2 Design-Based Small Area Estimation: The Nearest Neighbour

In practice, most large-scale sample surveys have complex designs, including multi-stage and multi-phase probability sampling procedures with stratification and clustering. Sampling weights, expansion weights, are calculated as the inverse of the overall inclusion probability with possible adjustments for non-response and other calibration factors. Let S be a sample selected with a complex survey design, let $Y_i, i \in S$ be the sample observations made on the variable of interest Y, and let $w_i, i \in S$ be a set of expansion weights. The population total and mean can be estimated as:

$$\hat{T}_w = \sum_{i \in S} w_i Y_i, \qquad \hat{M}_w = \sum_{i \in S} w_i Y_i \,/\, \sum_{i \in S} w_i \tag{1}$$

The variance estimation can be calculated approximately by Taylor Linearization approximation, or by the Jackknife repeated replication method.

Suppose that the total population can be subdivided into small areas U_a, $U = \bigcup_1^A U_a$, with unknown small area totals $T_a = \sum_{i \in U_a} Y_i$, $1 \le a \le A$. Suppose that the sample S can be subdivided the same way into small sub-samples S_a, $S = \bigcup_1^A S_a$ with a small sample size for each area. The aim is to efficiently estimate the area total or its mean based on S_a for each small area. Direct estimation based only on a small area sample is usually inefficient because of the small sample size. There are many ways to construct small area estimates by borrowing 'strength' based on spatial or structural properties of the small area, including design-based, model-assisted, and model-based methods [1-5]. Design-based small area estimation techniques use auxiliary information from outside the survey data to improve the reliability of the direct estimates, including ratio estimator, regression estimator, or more generally, calibration estimators.

In this study, we explore a design-based small area estimation method that uses the nearest neighbour technique. Sampling units located geographically close or close in other related measures correlated with the variable of interest may have similar characteristics to the study variables. We pool the sampled sampling units "close" to the small area together with the sampled sampling units from the small area to form a group, a nearest neighbourhood, and then treat it as a survey domain. A domain created in this way could be a true survey domain or a pseudo domain, depending on the definition of the distance measure. If the distance measure defines a fixed sub-population that does not depend on any sample selection results, then the domain is a true survey domain. For example, all sampling units located within a fixed distance from a fixed geographical point within a small area form a true survey domain. All sampling units located within a fixed distance to any of the sampled sampling units of a small area form a true survey domain. By treating them as survey domain, we can use all the known statistical inference techniques of survey domain analysis, including the estimation of small area totals, means and their variance estimation. Let S_a^+ be the enlarged sample consisting of the small area sample plus the borrowed sampling units from nearest neighbour areas; then the small area total can be estimated by

$$\hat{T}_a^* = N_a \times \frac{\sum_{i\in S_a^+} w_i Y_i}{\sum_{i\in S_a^+} w_i} \quad \text{or} \quad \hat{T}_a^* = \hat{N}_a \times \frac{\sum_{i\in S_a^+} w_i Y_i}{\sum_{i\in S_a^+} w_i} \tag{2}$$

depending on whether or not the small area population size N_a is known, where $\hat{N}_a = \sum_{i\in S_a} w_i$ is the estimate of the area population size in case it is unknown.

When similar surveys have been conducted in the same area in recent years and the characteristics to be estimated are stable over time, we can then combine two surveys together to increase the sample size for small areas. Let $S_a^{(1)}$ and $S_a^{(2)}$ be the small area samples from the previous survey and the current (target) survey, respectively, and $\hat{T}_a^{(1)}$and $\hat{T}_a^{(2)}$ are the direct estimates of area total over small area a, the following estimate

$$\hat{T}_a^* = \alpha\hat{T}_a^{(1)} + (1-\alpha)\hat{T}_a^{(2)} = \alpha \sum_{i\in S_a^{(1)}} w_i\, Y_i + (1-\alpha) \sum_{i\in S_a^{(2)}} w_i\, Y_i \tag{3}$$

is an estimate for the current area total through a proper weighting factor α $(\alpha > 0)$. The two direct estimates can be weighted equally or weighted with an importance weight. Assuming that the two surveys are independent, then all analysis is simple and direct based on standard survey data analysis tools and techniques. We use simple notations in the formula, but the values of Y_i in the two different terms represent the sample values of the variable of interest at different occasions.

It is desirable that small area estimates that are produced by different methods are consistent with reliable higher-level estimates. The small area estimates may be adjusted for consistency. Let $\hat{T}_a^{\Delta}$ be the adjusted estimate such that

$$\hat{T}_B = \sum_{a\in B} \delta_a \hat{T}_a^{\Delta}, \qquad \delta_a = \frac{\sum_{i\in S_a} w_i}{\sum_{i\in S_B} w_i} \tag{4}$$

where $\hat{T}_B$ is the broad area or higher-level estimate based on the full sample S_B from the broad area B, $\hat{T}_B = \sum_{i\in S_B} w_i\, Y_i$ .The simplest adjustment is

$$\hat{T}_a^{\Delta} = \frac{\hat{T}_B}{\sum_{a\in B} \delta_a \hat{T}_a^{*}} \times \hat{T}_a^{*} \tag{5}$$

$\hat{T}_a^{\Delta}$ is consistent in the sense that it can be aggregated to the broad area estimate $\hat{T}_B$. A more complex adjustment uses a "reverse calibration" procedure by treating $\hat{T}_B$ as the target total and the $\hat{T}_a^{*}$s as "weights". This adjustment can also be used to adjust complex indicators such TFR and CMR.

The proposed methods in this study are different from the Broad Area Ratio Estimator (BARE) ADB (2020) which pools all neighbouring small areas together from a broad area, where a homogeneous assumption was made that all small areas have the same mean as the broad area. It is also different from the reweighting method of Schirm and Zalansky et al. (1997) which uses the full sample including the small area and adjusts the sampling weights to catch the small area population size or other known population characteristics.

In the following sub-sections, we use data from the Rwanda DHS 2014 as the target survey and Rwanda DHS 2010 as the auxiliary survey. Rwanda has five provinces, each of which is subdivided into districts, with a total of 30 districts. The two surveys had the same design as two-stage cluster surveys, both of which had a designed sample size of 492 clusters and 12,792 households, and 26 households per cluster. The sample size was 16 clusters and 416 households per district, except for the three districts in Kigali province where 20 clusters and 520 households per district were sampled. The district level sample size is not too small for many indicators, but it is too small for a direct estimate of TFR. What we call "small area" here is relative to specific variables where a reliable estimate usually needs a very large sample size.

2.2.1 The Time-Space Nearest Neighbour

This method uses time-space nearest neighbour by simply combining the 2010 and 2014 surveys together, since all the sampled clusters in the 2010 survey in a district are the nearest neighbours geographically and maybe in time for the clusters in the same district for the 2014 survey. This doubles the sample size for each of the 30 districts and meets the minimum sample size requirement for TFR estimation at the domain level. All districts have 32 clusters and 984 households, except three districts in Kigali province each have 40 clusters and 1040 households. TFR is calculated using a standard procedure based on the combined sample as they were from a single survey. However, an importance weight can be used to reflect the user's objective judgement. For example, a larger weight can be assigned to the target survey than to the auxiliary survey. The importance weight can be area/district specific. In this study, we used equal weights and province level TFR variance weights for simplicity, but two different weights give very close results. The results reported here used the equal weight option; the calculated TFR without adjustment represents a reference period between the two surveys, roughly from 2010 to 2011. A consistency adjustment for the 2014 provincial TFR makes the estimate lean toward the 2014 survey. A consistence adjustment can be made for age-specific fertility rates or the total fertility rate. The results are adjusted for TFR at the province level.

2.2.2 The District Centre Nearest Neighbour

This method uses the small area centre point as a reference point and calculates the distance of the other clusters from other areas and takes a few clusters closest to the small area central point as the nearest neighbourhood, which creates a true survey domain. Usually, the small area central point is an easy-to-get information. When the sample size from the small area is not too small, such as for the Rwanda DHS 2014, a central point calculated based on the GPS coordinates of the sample clusters should be very close to the district central point, the central point of inhabited areas, which is better and more meaningful than the actual geographical centre. Suppose a group of such clusters are identified, plus the sample from the target district, the population characteristic estimation has the same formula as equation (2). The 2014 Rwanda DHS collected the GPS central point for each of the 492 clusters. We calculated the district centre based on the sample points from the district, and then calculated the distance to a district centre for each of the clusters which are not from the target district and took the first 20 clusters closest to the targeted district centre. Some districts borrowed clusters only from other districts in same province, and some districts borrowed clusters from districts in other provinces.

2.2.3 The Cluster Centre Nearest Neighbour

This method uses the cluster centre's nearest neighbour. Distances to a cluster in the target district from each of the clusters in neighbouring districts are calculated from the same province or other provinces. The 5 clusters closest to each cluster in the target district are identified. Since one neighbour cluster can be the nearest neighbour for several clusters in the target district, we just kept the distinct clusters. A fixed distance cut-off was used to control the neighbourhood size. The distance cut-off was district specific, ranging from 4 km to 25 km, with a target of about 16 distinct clusters borrowed for each district to reach the smallest sample size required for reliable TFR estimation at the district level. The number of borrowed clusters ranges from 8-19 per district. Some districts borrowed clusters only from the neighbouring districts within the same province, and some districts also borrowed clusters from districts in other provinces.

3 Numerical results

In this section, we present numerical results using the Rwanda DHS 2010 and Rwanda DHS 2014 data and the SAE methods proposed in the previous sections to estimate district level TFR. TFR is the average number of children a woman would have by age 49 if she bore children at current age-specific fertility rates. It is calculated as the sum of 7 age-specific fertility rates (ASFR) by 5-year age group multiplied by 5 for the 3 years before the survey. An age-specific fertility rate is the ratio of the number of live births over women-years of exposure. We calculated the direct estimates and the three SAE estimates; their variance and confidence intervals were calculated using the Jackknife method.

Confidence intervals for direct estimates have an average length of 1.43 children, which is beyond our precision control for domain level estimation which is controlled for one child. The average confidence interval length for the three SAE estimates is 1.01 (SAE1) for the combined estimate, 0.93 (SAE2) for the district centre nearest neighbour method, and 0.96 (SAE3) for the cluster centre nearest neighbour method. These are all under our controlled precision for domain level estimation. The two nearest neighbour methods produce very similar results. Figure 1 shows the numerical results, plotted against the provincial level TFR. The curve of the direct estimate departs more from the provincial level curve, SAE1 is better, and SAE2 and SAE3 are very close to the provincial level curve.

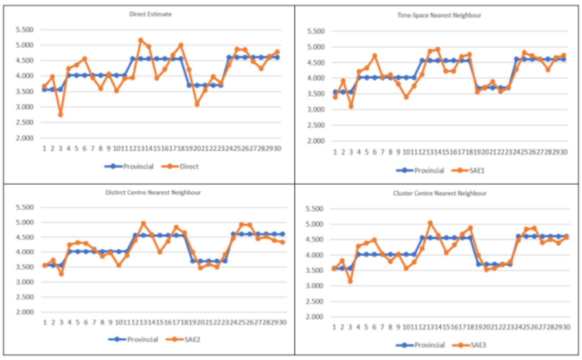


Figure 1. Line plot of the various TFR estimates against the provincial TFR estimates

References

1. Asian Development Bank (2020). *Introduction to small area estimation techniques: a practical guide for national statistics offices.*
2. Ghosh, M & Rao, J N K (1994). Small area estimation: an appraisal. Statistical Science. Vol. 9, pp 55-93.
3. Marker, D A (1999). Organization of small area estimators using a generalized linear regression framework. Journal of Official Statistics. Vol. 15, pp 1-24.
4. Särndal, C E (1984). Design-consistent versus model-dependent estimation for small domains. JASA, Vol. 79, pp 624-631.
5. Schirm A L, Zaslavsky, A M & Czajka L (1997). Large numbers of estimates for small areas. *Mathematica Policy Research Report.*

Small area estimation via multivariate generalized linear mixed effects models

E. Rocco [a*] and M.F. Marino [b*]

[*]Department of Statistics, Computer Science, Applications, University of Florence

Abstract

Analysing complex phenomena often requires the estimation of multivariate, correlated, descriptive parameters, which may potentially have a heterogeneous nature: binary variables, counts, continuous symmetric or skewed variables, or a combination of them. Moreover, a frequent issue is that of deriving estimates for small spatial areas that are non-sampled or are under-sampled in surveys. Estimates of such parameters can 'borrow strength' from data on multiple characteristics and/or auxiliary variables from other neighboring areas through appropriate models. We suggest the use of a multivariate mixed effect model, based on correlated random effects, for the jointly modelling of multiple outcomes recorded on a sample of units clustered within small areas. This allows us to account for the multivariate dependence among outcomes by means of the latent terms in the model. The proposal is tested by means of an intensive simulation study considering different types of outcomes.

Keywords— multiple characteristics, multivariate unit-level small area models, correlated random effects

1 Introduction

Direct small area estimates (SAE) may not provide acceptable precision in the presence of small sample sizes or out-of sample areas. In this framework, indirect, model-based, approaches represent a powerful tool since they allow to borrow information across related areas by means of auxiliary variables. The most popular models for SAE are the linear-mixed models. They area based on independent, area-specific, random effects that allow us to account for the variability between areas exceeding that explained by the auxiliary variables. Responses can be either observed at the small area-level or at a the unit-level. Fay and Herriot (1979) studied the area-level model and proposed

[a]emilia.rocco@unifi.it

[b]mariafrancesca.marino@unifi.it

an empirical Bayes estimator for this case. Battese and Fuller (1988) considered the unit-level model and constructed an empirical best linear unbiased predictor (EBLUP) for the small area means. Several extensions to this set-up have been considered in the literature, mostly for handling univariate survey data. Rao and Molina (2015) provides a general reviews of small area estimation. When the aim is to estimate a finite population mean vector of multiple characteristics, multivariate area-level or unit-level mixed models allow to take into account their correlation. For multivariate area-level data, Fay (1987) proposed the multivariate Fay-Herriot model; some extensions have also been considered in the literature (e.g., Porter et al., 2015; Ubaidillah et al., 2019). For multivariate unit-level data, the use of multivariate linear mixed models has been considered by Datta et al. (1999) and Esteban et al. (2020) among others. All the cited works prove how the multivariate approach allows to achieve substantial improvement over the univariate counterparts. However, in the SAE literature, the use of multivariate generalized linear mixed models (GLMMs) to estimate a finite population mean vector of non-Gaussian, multiple, characteristics observed at the unit-level has not been studied much. In this paper, we suggest and investigate its use to deal with the estimation of a mean vector for each small area when the assumption of the multiple linear mixed model are not satisfied. Section 2 describes the estimation setting while Section 3 carries out simulation experiments and reports a final discussion.

2 Small area estimation under multivariate GLMMs

Let U be a finite population of N units, partitioned in m small-areas ($U_i \subset U$) of size N_i, with $\sum_{i=1}^{m} N_i = N$. Let $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijD})'$ denote an D-dimensional response vector for unit j belonging to small area i ($i = 1, \ldots, m, j = 1, \ldots, N_i$). Let $u_i = (u_{i1}, \ldots, u_{iD})'$ be an area-specific random vector having response-specific components. Last, let x_{ij} denote a p-dimensional vector of auxiliary covariates, and X_{id} the corresponding matrix of covariates for the i-th small area and the d-th response.

We assume that the following generalized linear mixed model relates the response variables in Y_{ij} to the auxiliary ones

$$\begin{cases} g_1(E[Y_{ij1} \mid u_{i1}]) = x'_{ij1}\beta_1 + u_{i1} \\ g_2(E[Y_{ij2} \mid u_{i2}]) = x'_{ij2}\beta_2 + u_{i2} \\ \vdots \\ g_D(E[Y_{ijD} \mid u_{iD}]) = x'_{ijD}\beta_D + u_{iD}], \end{cases} \tag{1}$$

Here, $g_d(\cdot)$ is a proper link function, x_{ijd} is a subset of x_{ij} (that is, $x_{ijd} \subseteq x_{ij}$), β_d is a p-dimensional vector of fixed model parameters describing the effect of covariates x_{ijd} on the (transformed) mean of Y_{ijd}, and u_{id} is the area- and response-specific random effect which is meant to describe sources of unobserved heterogeneity not captured by x_{ijd}. We assume that each element in Y_{ij}, say Y_{ijd}, is influenced by the corresponding element in u_i only, that is u_{id}, and conditional on these effects, multiple responses from the same unit are independent (within-unit local independence). The proposed multivariate small area model is completed by the following assumptions.

- Conditional on the vector u_i, responses from units in the i-th area, $Y_i = \{Y_{i1}, \dots, Y_{iN_i}\}$, are independent (within-area local independence) and the corresponsing joint, conditional, density is:
$$f(y_i \mid u_i) = \prod_{j=1}^{N_i} f(y_{ij} \mid u_i) = \prod_{j=1}^{N_i} \prod_{d=1}^{D} f(y_{ijr} \mid u_{id})$$
where $f(y_{ijd} \mid u_{id})$ denotes the Exponential Family (EF) density.
- The area-specific random vector u_i follows a zero mean, multivariate, Gaussian distribution, with unconstrained covariance matrix Σ_u. Diagonal elements of such a matrix identify the variance of the area- and response-specific random effect $u_{id}, d = 1, \dots, D$, while off-diagonal elements correspond to the covariance between couples $(u_{id}, u_{id'})$. This latter provides an (indirect) measure of dependence between the corresponding responses $(Y_{ijd}, Y_{ijd'})$.

Our aim is here that of predicting the vector of small area means $\bar{y}_i = (\bar{y}_{i1}, \dots, \bar{y}_{iD})'$ with $\bar{y}_{id} = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ijd}$, $d = 1, \dots, D$, by using the GLMM in equation (1). To this aim, a sample (s_i) of n_i units is selected from each small area U_i according to a non-informative sampling design ($\cup_{i=1}^{m} s_i = s$). We assume that covariates X_{id} are available at the population level, while responses Y_{ij} are observed for sampled units only. In this respect, each component in $\bar{y}_i$ can be partitioned into the sampled and unsampled part
$$\bar{y}_{id} = \frac{1}{N_i} \sum_{j \in s_i} y_{ijd} + \sum_{j \notin s_i} y_{ijd},$$
and predictions for y_{ijd} when $j \notin s_i$ can be obtained as
$$\hat{y}_{ijd} = g^{-1}(\hat{u}_{id} + x'_{ijd} \hat{\beta}_d).$$
Here, $\hat{\beta}_d$ and $\hat{u}_{id}$ denote the maximum likelihood estimate of fixed model parameters and the empirical best predition of the area- and response-specific random effectin the model, respectively. The latter is obtained as $\hat{u}_{id} = E(u_{id} \mid y_{id})$.

While the prediction of one small-area mean at a time from univariate equations in (1) allows us to obtain equivalent results to the proposed multivariate approach in terms of bias, improvements are expected to be observed in terms of efficiency. Indeed, as far as this measure is entailed, using the multivariate approach allows us to borrow strength not only from areas (as for the univariate approach), but also from multiple responses. Furthermore, the proposed multivariate small area model directly nests the corresponding univariate ones. When responses are uncorrelated, the covariance matrix for the area-specific random effects reduces to $\Sigma_u = \sigma_u I_D$, where I_D denotes a $D \times D$ identity matrix. Last, but not least, it can be the case that analysing the association structure between multiple responses is itself of interest. In this sense, the proposed multivariate approach represents an iteresting tool of analysis which is worth to consider.

3 Simulations

In order to investigate the performance of the proposed multivariate SAE approach, a large scale model-based simulation experiment has been performed. It takes into account several scenarios. For each of them, bivariate population data are generated from the following bivariate GLMM:

$$\begin{cases} g_1(E[Y_{ij1} \mid u_{i1}]) = \beta_0 + x_{ij}\beta_{11} + u_{i1} \\ g_2(E[Y_{ij2} \mid u_{i2}]) = \beta_0 + x_{ij}\beta_{12} + u_{i2}, \end{cases}$$

where $g_k(\cdot), k = 1,2$, denotes a proper link function, a single covariate x is considered, and the area-specific effects $\mathbf{u}_i = (u_{i1}, u_{i2})'$ are simulated from a bivariate Gaussian distribution $\mathbf{u}_i \sim N_2(\mathbf{0}, \Sigma_u)$ with two specification for the covariance matrix

$$\Sigma_u^{(high)} = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}, \qquad \Sigma_u^{(low)} = \begin{bmatrix} 1 & 0.32 \\ 0.32 & 1 \end{bmatrix}.$$

Concerning the response type, the related link function, and the values of fixed model parameters, we have considered the following two settings:

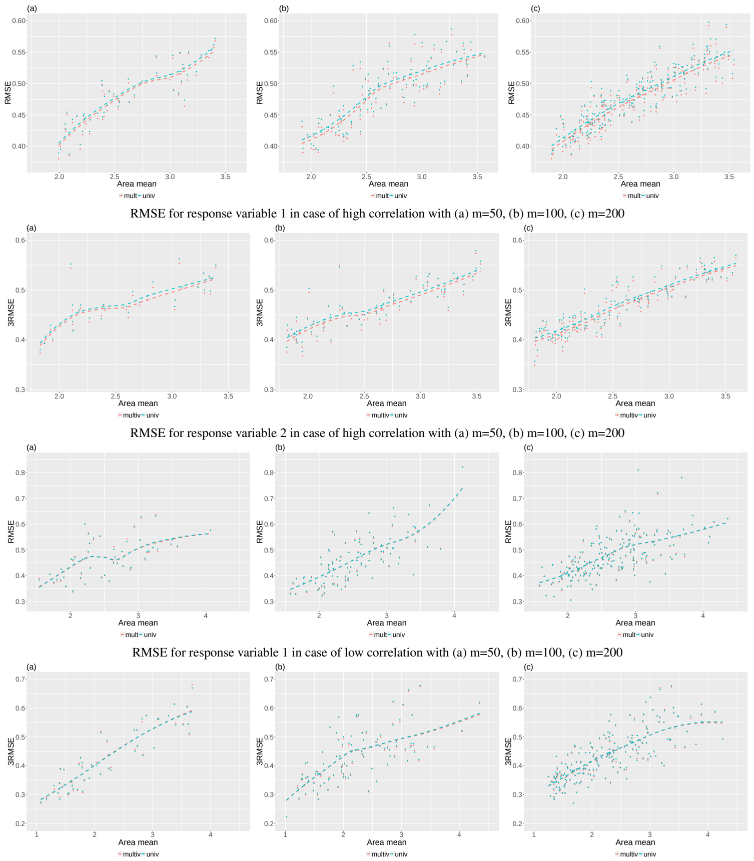
1. A pair of Poisson responses – $Y_{ij1} \mid u_{i1} \sim Pois(\cdot)$ and $Y_{ij2} \mid u_{i1} \sim Pois(\cdot)$ – with $g_1(\cdot) = g_2(\cdot) = \log(\cdot)$ and regression parameters: $\beta_0 = 0.7, \beta_{11} = -0.1, \beta_{12} = -0.2$

2. A pair of Bernoulli responses – $Y_{ij1} \mid u_{i1} \sim Bern(\cdot)$ and $Y_{ij2} \mid u_{i1} \sim Bern(\cdot)$ – with $g_1(\cdot) = g_2(\cdot) = \text{logit}(\cdot)$ and regression parameters: $\beta_0 = 0.5, \beta_{11} = -0.4, \beta_{12} = -0.6$

The population size is assumed to be constant across areas ($N_i = 100$ for $i = 1,...,m$), while three different values are considered for the number of areas: $m =$ 50, 100, 200. For each m, the auxiliary variable in the model is generated as $x_{ij} \sim \text{Unif}(1, i/b)$, where $b = 4, 8, 16$, for $m =$ 50, 100, 200, respectively.

For each scenario (for a total of $2 \times 2 \times 3$), sample data are drawn by using a stratified sampling design, with strata corresponding to areas and equal sample size, $n_i = 10$, in each stratum. The number of simulation runs is fixed at $T = 1000$. Finally, the multivariate small area estimates obtained through the proposed multivariate GLMM are compared to estimates obtained through the corresponding univariate coutnerparts. To this end, for each area i and each predictor, the empirical Root Mean Squared Error (RMSE) is calculated as: $RMSE_i = \sqrt{T^{-1} \sum_{t=1}^{T} (\hat{\bar{y}}_{it}^{Model} - \bar{y}_{it})^2}$.

From the empirical results shown in Figure 1 and 2, it is evident that in the presence of highly correlated responses, the multivariate modelling is preferable to the univariate counterparts, whatever the nature (the distribution form) of the data. When the correlation is low, may be opportune to evaluate for each case (considering the nature of the data, the number of small-areas, the size of the sample, and the aim of the study) the trade of between the capability of the multivariate approach to exploit the relation among the two variables and the higher complexity of the model itself. However, the problem in the case of low correlation may be not the mean predictor itself, but the corresponding MSE estimators based on an incorrect model. When the target variables are positively correlated, univariate models usually tend to under-estimate the MSEs. Prospective research endeavors will investigate more this aspect and will consider the extension to multivariate mixed type responses.



RMSE for response variable 1 in case of high correlation with (a) m=50, (b) m=100, (c) m=200

RMSE for response variable 2 in case of high correlation with (a) m=50, (b) m=100, (c) m=200

RMSE for response variable 1 in case of low correlation with (a) m=50, (b) m=100, (c) m=200

RMSE for response variable 2 in case of low correlation with (a) m=50, (b) m=100, (c) m=200

Figure 1: Poisson data

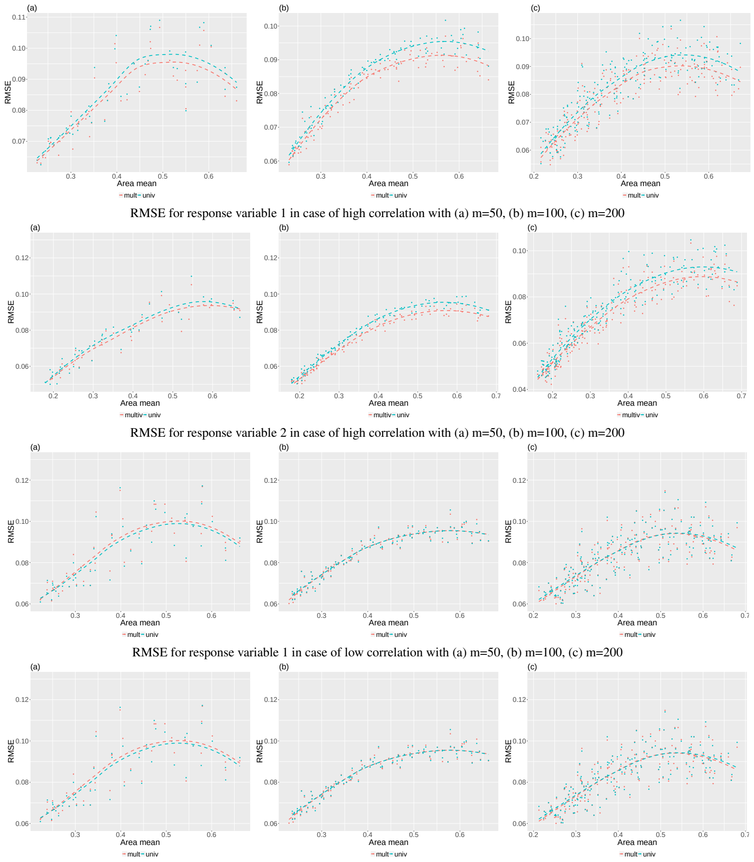
References

Battese, G. E. Harter, R. and Fuller, W. A. (1988). An error component model for prediction of county crop areas using survey and satellite. *Journal of the American Statistical Association*, 83:28–36.

Datta, G. S., Day, B., and Basawa, I. (1999). Empirical best linear unbiased and empirical bayes prediction in multivariate small area estimation. *Journal of Statistical Planning Inference*, 75:264–279.

Esteban, M. D., Lombardía, M. J., López-Vizcanío, E., Morales, D., and Pérez, A. (2020). Small area estimation of expenditure means and ratios under a unit-level bivariate linear mixed model. *Journal of Applied Statistics*.

Fay, R. E. (1987). Application of multivariate regression of small domain estimation.



RMSE for response variable 1 in case of high correlation with (a) m=50, (b) m=100, (c) m=200

RMSE for response variable 2 in case of high correlation with (a) m=50, (b) m=100, (c) m=200

RMSE for response variable 1 in case of low correlation with (a) m=50, (b) m=100, (c) m=200

RMSE for response variable 2 in case of low correlation with (a) m=50, (b) m=100, (c) m=200

Figure 2: Bernoulli data

In Platek, R., Rao, J. N. K., Särndal, C. E., and Singh, M. P., editors, *Small Area Statistics*, pages 91–102. New York: John Wiley.

Fay, R. E. and Herriot, R. A. (1979). Estimation of income from small places: an application of james-stein procedures o census data. *Journal of the American Statistical Association*, 74:269–277.

Porter, A. T., Wikle, C. K., and Holan, S. H. (2015). Small area estimation via multivariate fay-herriot models with latent spatial dependence. *Australian & New Zealand Journal of Statistics*, 57:15–29.

Rao, J. and Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons, Ltd, Hoboken, New York, 2 edition.

Ubaidillah, A., Notodiputro, K. A., Kurnia, A., and Wayan, I. (2019). Multivariate fay-herriot models for small area estimation with application to household consumption per capita expenditure in indonesia. *Journal of Applied Statistics*, 46:2845–2861.

Small area estimation via Heteroskedastic Geographically Weighted Regression for functional data

Elvira Romano [a*], Andrea Diana [b*] and Jorge Mateu [c**]

[*]Department of Mathematics and Physics, Universitá della Campania Luigi Vanvitelli, Caserta, Italy
[**] Department of Mathematics, Campus Riu Sec, University Jaume I, Castellon, Spain

Abstract

Small area estimation is studied under a Heteroskedastic Geographically Weighted Regression model for functional data. The calibrated spatio-functional model we propose assumes that the variance varies across the space, and that each local model (defined at each location) gives a local non parametric estimation of the variance. This approach improves the model performance in terms of predictive spatio-functional fit for small area estimation, as illustrated by a simulation study and financial data analysis.

Keywords— Small are estimation, Functional data analysis, Heteroskedasticity, Non-stationarity, Weighted regression

1 Introduction

Nowadays, many modern systems need the estimation and publication of statistics for disaggregated domains (Marchetti et al., 2015). This is strictly related to small area estimation problems involving the estimation of parameters for small sub-populations, generally defined when the sub-population of interest is included in a larger survey.

In this framework robust methods like *GWR* proposed by Salvati et al. (2012) have proven to be very useful. In this paper we propose a robust small area methodology that allows for the presence of spatial correlation in the data. We especially present a robust predictive functional approach that incorporates the spatial impact from the areas on the small area of interest.

[a] elvira.romano@unicampania.it
[b] andrea.diana@unicampania.it
[c] mateu@mat.uji.es

The approach, we present, is a generalisation of a geographically weighted regression model (GWR) (Fortigam et al., 2002) for spatially correlated sample curves, focusing on the problem of non-stationarity in parameter estimation. Like in (Romano et al., 2020), we calibrate the variance of the model by estimating locally the variance and we improve the model performance in terms of predictive fit by searching a parameter-specific distance metric. This involves replacing the weighted linear regression in IRLS algorithm by a back-fitting algorithm.

2 Heteroskedastic GWR model (H-GWR) for spatially dependent functional data

Let $Y_D = \{Y_s(t) : t \in T,\ s \in D\}$, be a functional response variable (Ramsay et al., 2005) observed at a location $s \in D \subseteq \mathbb{R}^2$, whose realisation as a function of $t \in T$ is a functional data, where T is a compact subset of $\mathbb{R}$. Assume we also have K functional covariates defined on T (Delicado et al., 2010), and observe $\boldsymbol{\chi}_D(v) = [\chi_{D,1}(v), \dots, \chi_{D,K}(v)]^T$. We introduce a model calibration by means of local estimation of the squared residuals. The model is built on the assumption that the variance varies across the space, and that each local model (defined at each location) gives a local estimation of the variance.

A GWR model can be defined written as

$$Y_s = m_s + \varepsilon_s \tag{1}$$

where m_s is a drift term and ε_s is a zero mean, second-order stationary and isotropic random field, so that

(i) $\mathbb{E}(Y_s) = m_s\ \forall s \in D$

(ii) $\mathbb{E}(\varepsilon_s) = 0\ \forall s \in D$

(iii) $\mathbb{V}ar(\varepsilon_s) = \mathbb{E}[\varepsilon_s \varepsilon_s^T] = \Sigma$, where the diagonal elements of Σ reflect zero spatial autocorrelation, and are defined by $\sigma^2(t) = Var(\varepsilon_s(t))$ that are independent from the spatial location.

At the generic location s_i, the model in (1) becomes

$$Y_{s_i}(t) = m_{s_i}(t) + \varepsilon_{s_i}(t) \quad i = 1, \dots, n \tag{2}$$

where $m_{s_i}(t)$ can be written as $m_{s_i}(t) = \beta_{s_i,k}(t) + \sum_{k=1}^{K} \int_T \chi_{s_i,k}(v) \beta_{s_i,k}(v,t,s_i) dv$, where $\beta_{s_i,0}$ is the intercept, and $\beta_{s_i,k}(v,t,s_i)$ are the functional coefficients of the k covariates at site s_i. Recall that the GWR prediction variance at a generic location s_i, without any assumption of spatial dependence, is defined as

$$\sigma^{2\,GWR}_{s_i}(t) = var\{\hat{Y}_{s_i}(t) - Y_{s_i}(t)\} = \hat{\sigma}^2(t)[1 + S_{s_i}(t)] \tag{3}$$

where

- $\hat{\sigma}^2(t) = RSS(t)/(n - ENP)$, with $RSS(t)$ the residual sum of squares and ENP the effective number of parameters of GWR fit. This is a function independent from the spatial location.

- $S_{s_i}(t)$ is the i-th element of the matrix $\mathbf{S} = (\mathbf{C}_k J_{\phi_k})\mathbf{W}_{s_i}(\sum_{k=1}^{K} \mathbf{C}_k J_{\phi_k} \mathbf{B}_{s_i,k}) J_{\phi_k}$.

In several cases, like in the small area estimation, the above assumption of independence of the variance from the space is not realistic. Indeed, the local variability of the coefficients in space may easily depend on different levels of the spatial variability. Thus to predict the function Y_{s_i} taking into account the non-constant spatial variability, we propose to correct the variance of the model providing local error variance estimates that are spatially dependent.

We calibrate the variance of the model ${\sigma^2}_{s_i}^{GWR}(t)$ by replacing $\hat{\sigma}^2(t)$ with $\hat{\sigma}^2_{s_i}(t)$, and assuming the latter is a continuous function over the space, we can estimate it by a mean smoother. The final variance $\hat{\sigma}^2_{s_i}(t)$ replaces $\hat{\sigma}^2(t)$ to give

$${\sigma^2}_{s_i}^{GWR}(t) = var\{\hat{Y}_{s_i}(t) - Y_{s_i}(t)\} = \hat{\sigma}^2_{s_i}(t)[1 + S_{s_i}(t)] \tag{4}$$

Note that in this equation, the random variable is ${\sigma^2}_{s_i}^{GWR}(t)$. For the local variance estimation, we need to model the relationship with the local mean. Thus we write the local mean in terms of a local smoother as follows

$$m_{s_i}(t) = \sum_{j=1}^{n} w_{s_i,s_j} Y_{s_i}(t) / \sum_{j=1}^{n} w_{s_i,s_j} = \sum_{j=1}^{n}\sum_{l=1}^{L} w_{s_i,s_j} a_l(t) f_l(s_i) / \sum_{j=1}^{n} w_{s_i,s_j} \;\; s_i, s_j \in D, t \in T, \tag{5}$$

where $f_l(\cdot)$, $l = 1, \dots, L$ are known functions at each location, and $a_l(\cdot)$, $l = 1, \dots, L$ are unknown functional coefficients independent of the spatial location that have to be estimated. Note that the dependence of the mean on the spatial location comes from the function $\{f_l(\cdot)\}_{l=1,\dots,L}$ and the weights w_{s_i,s_j}.

Then, the local variance smoother becomes

$$L_{\sigma^2_{s_i}}(t) = \sum_{j=1}^{n} w_{s_i,s_j}(Y_{s_i}(t) - m_{s_i}(t))^2 / \sum_{j=1}^{n} w_{s_i,s_j}. \tag{6}$$

This is a mean smoothing over the observed squared residuals, and provides the following local variance estimation

$$\hat{\sigma}^2_{s_i}(t) = \sum_{j=1}^{n} w_{s_i,s_j}(Y_{s_i}(t) - \sum_{j=1}^{n}\sum_{l=1}^{L} w_{s_i,s_j} a_l(t) f_l(s_i) / \sum_{j=1}^{n} w_{s_i,s_j})^2 / \sum_{j=1}^{n} w_{s_i,s_j}.$$

Spatially varying relationships between the dependent variable and the covariates are accounted for by locally weighting and calibrating from the spatial locations. The algorithm is iterated with updated estimates of $\hat{\beta}_{s_i,k}(v,t,s_i)$ until an acceptable level of convergence is reached. Together with the parameter estimates, the H-GWR prediction at s_i is also updated. In a traditional GWR technique, the calibration given by $\mathbf{W}_{s_i}$ is obtained by kernel smoothing where the bandwidth is selected via a leave-one-out cross-validation method or by the Akaike Information criterion (AIC). In both cases the Euclidean distance is often used as a default metric in this calibration step. In addition to a first heteroskedastic calibration step, we additionally propose to account for the spatial non-stationarity in the parameter estimates by building a weight matrix using a back-fitting algorithm, generalization of (Liu et al., 2015).

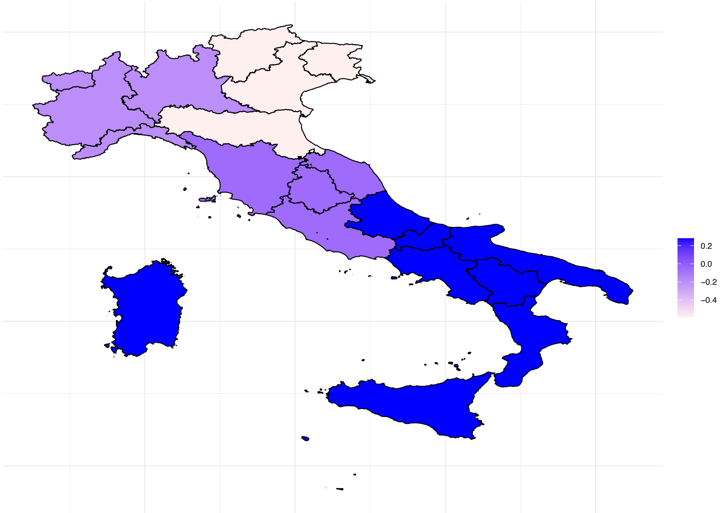


Figure 1: Distribution of the difference between residuals of two estimated models $HGWR$ calibrated and not calibrated.

3 Application results

Banks have always been the main source of finance for the Italian economy, and learning about the strengths and weaknesses of the banking system is essential to understand the economic prospects of the country, in particular given the growing integration of international financial markets. The relationship between financial systems and growth has been explored at the infra-national level as local finance and growth stream of studies. In this perspective, the aim of this study is to show the cross-regional variation in economic growth in countries such as Italy by using the H-GWR technique, proposed in Section 2. Especially the impact of years (2000-2014) of transformations of the Italian banking system on the local economic development has been investigated.
Data (collected from the Bank of Italy database (www.bancaditalia.it)) consists of a panel composed by the 103 Italian provinces, considering for each province the time series over 15 years of the following variables: the value added by worker ($Y_D(t)$); the percentage of the ratio between the number of Banche di Credito Cooperativo (BCC) and number of total bank ($X_{D,1}(\tau)$); the ratio of total loans on total value added ($X_{D,2}(\tau)$).

We considered a functional model to estimate $Y_D(t))$ from $X_{D,1}(\tau)$ and $X_{D,2}(\tau)$. In particular, in the aftermath of the $2007-08$ global banking crisis, Italy underwent a credit crunch that particularly affected small, local cooperative banks. During a credit crunch such small banks may be more inclined to reduce lending to their traditional clientele. This may weaken or cancel the negative effect of reduced banking diversity on growth at the local level.
We considered a functional model to estimate $Y_D(t))$ on $X_{D,1}(\tau)$ and $X_{D,2}(\tau)$ over the

period $2000-2014$.

The 103 Italian provinces divided in 4 areas: Sud, Midlle, Nord-Est, Nord-ovest. The mean of residuals was calculated for each area, with HGWR calibrated method (mrC_{HGWR}) and HGWR method (mrC_{HGWR}). Figure 3 shows a map of Italy painted different level of blue. Light areas are the macro areas where the means of residuals of the calibrated HGWR are smaller than the means of residuals of the HGWR.

References

Delicado, P., Giraldo, R., Comas, C. and Mateu, J. (2010). *Statistics for spatial functional data: some recent contributions*. Environmetrics, 21, 224-239.

Fotheringham, A.S., Brunsdon, C. and Charlhon, M.E. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, New York.

Lu, B., Harris, P., Charlton, M.E. and Brunsdon, C. (2015). *Calibrating a geographically weighted regression model with parameter-specific distance metrics*. Procedia Environmental Sciences, 26, 109-114.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L. and Gabrielli, L. (2015). *Small Area Model-Based Estimators Using Big Data Sources*. Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 263281, http://dx.doi.org/10.1515/JOS-2015-0017

Ramsay, J.E. and Silverman, B.W. (2005). *Functional Data Analysis* (Second ed.). Springer.

Romano, E., Mateu, J., Butzbach O. (2020). *Heteroskedastic geographically weighted regression model for functional data*. Spatial Statistics, Volume 38.

Salvati,N., Tzavidis, N., Pratesi, M., Chambers, M. (2012). *Small area estimation via M-quantile geographically weighted regression*. Test 21 (1), 1-28.

Yamanishi, Y. and Tanaka, Y. (2003). *Geographically weighted functional multiple regression analysis: A numerical investigation.* Journal of Japanese Society of Computational Statistics, 15, 307-317.

Controlling the bias for M-quantile estimators for small area

Francesco Schirripa Spagnolo [a*], Gaia Bertarelli [b**], Raymond Chambers [c***], David Haziza [d****] and Nicola Salvati [e*]

[*]Dipartimento di Economia e Management, Università di Pisa, Pisa, Italy
[**]Istituto di Management, Scuola Universitaria Superiore Sant'Anna, Pisa, Italy
[***]National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, Australia
[****]Department of mathematics and statistics, University of Ottawa, Ottawa, Canada

Abstract

In this paper we propose two bias correction approaches in order to reduce the prediction bias of the robust M-quantile predictors in small area estimation in the presence of representative outliers. A Monte-Carlo simulation study is conducted. Results confirm that our approaches improve the efficiency and reduce the prediction bias of M-quantile predictors when the population contains units that may be influential if selected in the sample.

Keywords— Robust methods, Small Area Estimation, M-quantile

1 Introduction

Outliers can arise frequently in sample surveys, for instance regarding economic variables whose distribution are highly skewed the data distribution is highly skewed. Some outliers are sample elements whose data values are recorded incorrectly or are unique, consequently they can be can be somehow corrected or removed. However, other outliers may not associated with an error: the

[a]francesco.schirripa@ec.unipi.it
[b]gaia.bertarelli@santannapisa.it
[c]ray@uow.edu.au
[d]dhaziza@uottawa.ca
[e]nicola.salvati@unipi.it

sample values associated with these outliers have been correctly recorded and they cannot be considered as unique. According to (Chambers, 1986) they are ‘representative outliers’. Such outliers values are representative of the non-sampled part of the population and they can seriously affect the survey estimates. Consequently, several methods have been developed in order to mitigate the effects of outliers on survey estimates. The representative outliers are even more concerning in the small area estimation (SAE) context, where sample sizes are very small and the estimation is often model-based Chambers et al. (2014). Robust small area estimation has received considerable attention in last years. Among other, Chambers and Tzavidis (2006) propose a robust approach based on the M-quantile regression aiming at overcoming the issue of outliers by avoiding the normal assumption. Sinha and Rao (2009) addressed the same issue from the perspective of linear mixed models. However, these approaches use plug-in robust prediction replacing parameter estimates in optimal but outlier-sensitive predictors by outlier robust versions and they may introduce a prediction biases. Dongmo-Jiongo et al. (2013) and Chambers et al. (2014) proposed a bias correction method for models with continuous response variables. The main aim of this work is to propose general bias correction methods to reduce the prediction bias of the robust M-quantile predictors in SAE in the presence of outliers. Two approaches are studied. The first estimator is a unified approach to M-quantile predictors based on a full bias correction and it could be viewed as a generalization of Chambers (1986). The second proposal is developed following the conditional bias approach by Beaumont et al. (2013) and Dongmo-Jiongo et al. (2013).

2 Bias corrected M-quantile-based estimator

Let θ_i be a finite population parameter for area i. That is, θ_i is a well-defined function of the values of a random variable Y associated with the N_i elements of such a small area finite population of interest. For ease of notation, we assume that both Y and θ_i are scalar, and we denote

$$\theta_i = f(\mathbf{y}_{U_i}),$$

where $\mathbf{y}_{U_i}$ denotes the vector of population values of Y for small area i and f is a known function. A basic sample survey inference problem is then one of predicting the value of θ_i give a sample of $n < N$ values from $\mathbf{y}_U$. Without loss of generality we put $\mathbf{y}_s$ equal to the population sub-vector defined by these values, where s denotes the set of sampled population units. We define (i) $\mathbf{y}_{U_i}$ vector of population values of Y for area i with $U = \bigcup_{i=1}^{m} U_i$ with m is the number of small areas; (ii) $\mathbf{y}_{s_i}$ vector of sampled population values in small area i with $s = \bigcup_{i=1}^{m} s_i$. Suppose that, given $\mathbf{y}_{s_i}$ we can impute the remaining values $\hat{\mathbf{y}}_{U_i}$ denote this imputed vector. A popular method of predicting the unobserved value of θ_i is via the Plug-In Predictor (PIP)

$$\hat{\theta}_i = f(\hat{\mathbf{y}}_{U_i}). \tag{1}$$

Adopting a model-based approach, the empirical PIP for θ_i based on this plug-in approximation is

$$\hat{\theta}_i = f(\mathbf{y}_{s_i}, \{\hat{y}_{ij}^{opt}; j \in U_i - s_i\}) \tag{2}$$

where the set $U_i - s_i$ contains the $N_i - n_i$ indices of the non-sampled units, $\hat{y}_{ij}^{opt} = E[y_{ij}|\mathbf{y}_s; \boldsymbol{\delta} = \hat{\boldsymbol{\delta}}]$ is the plug-in approximation of the minimum mean squared error predictor (MMSEP) of y_{ij}^{opt} for a non-sampled population unit j for area i, and $\boldsymbol{\delta}$ is a vector of unkown parameters. The above PIP (2) for small area can be also computed using the M-quantile approach. It can be obtained by using the estimated regression coefficients by M-quantile approach, $\hat{\boldsymbol{\beta}}_{\tau}$, leading to

$$\hat{\theta}_i^{MQ} = f\Big(\mathbf{y}_{s_i}, \{g^{-1}(\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{\tau_i}); j \in U_i - s_i\}\Big), \tag{3}$$

where τ_i represents the order of M-quantile for area i. Its computation varies depending on the type of the data.

We propose two small area estimators based on Generalised version of M-quantile regression models.

The first estimator is a unified approach to M-quantile predictors based on a full bias correction. Following Chambers (1986), the first order approximation to the prediction bias of $\hat{\theta}_i^{MQ}$ is

$$E[\hat{\theta}_i^{MQ} - \theta_i] \simeq \sum_{j \notin s_i} \Big(\frac{\partial f}{\partial y_{ij}}\Big)_{\mathbf{y}_{U_i} = \mathbf{m}_{U_i}} E[\hat{y}_{ij} - y_{ij}] \simeq \sum_{i \in r_j} \Big(\frac{\partial f}{\partial y_{ij}}\Big)_{\mathbf{y}_U = \hat{\mathbf{m}}_{U\bar{q}_j}} \Big(\frac{\partial g^{-1}}{\partial \eta}\Big)_{\eta = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{\bar{q}_j}} \mathbf{x}_{ij}^T E[\hat{\beta}_q - \beta_q],$$

The bias corrected robust predictor MQC for the population average of Y in the ith area will be:

$$\theta_i^{MQC} = N_i^{-1} \left(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{\mu}_{ij} + \sum_{j \in r_i} \Big(\frac{\partial f}{\partial y_{ij}}\Big)_{\mathbf{y}_U = \hat{\mathbf{m}}_{U\bar{q}_j}} \Big(\frac{\partial g^{-1}}{\partial \eta}\Big)_{\eta = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{\bar{q}_j}} \mathbf{x}_{ij}^T \hat{\mathbf{B}}_i \right) \tag{4}$$

where $d_{jh\bar{q}_j} = 2\{\bar{q}_j I(r_{hj} > 0) + (1 - \bar{q}_j) I(r_{hj} \leq 0)\}$ and $\hat{\mathbf{B}}_i$ has to be computed depending of the type of the response variable. If y_{ij} is continuous

$$\hat{\mathbf{B}}_i = \left(\sum_{h=1}^{m} \sum_{j \in s_h} \mathbf{x}_{hj} \hat{d}_{hj} \mathbf{x}_{hj}^T \right)^{-1} \sum_{h=1}^{m} \sum_{j \in s_h} \mathbf{x}_{hj} \hat{d}_{hj} \hat{\sigma}_{hj} \phi \left\{ \frac{y_{hj} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{\tau_i}}{\hat{\sigma}_{hj}} \right\}. \tag{5}$$

The second proposal is developed following the conditional bias approach by Beaumont et al. (2013) and Dongmo-Jiongo et al. (2013). In a model based approach, the conditional bias attached to unit ij is

$$B_{ij} = E[\hat{\theta} - \theta | s; Y_{ij} = y_{ij}].$$

The prediction error $\hat{\theta}_i - \theta_i$ can be approximated as:

$$\hat{\theta}_i - \theta_i \simeq \sum_{j \in r_i} B_{ij}(I_{ij} = 0) + \sum_{j \in s_i} B_{ij}(I_{ij} = 1). \tag{6}$$

To determine the conditional bias, we need to distinguish two cases, whether the unit belongs to the sample or not. The main problem is that the conditional bias of a non-sampled unit can't be estimated since it depends on the Y-values on the non-sample units, which are not observed. A robust predictor of the mean in the ith area can be expressed as

$$N_i^{-1} \left(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} g^{-1}(\mathbf{x}_{ij}^T \boldsymbol{\beta}) - \sum_{j \in s_i} B_{ij}(I_{ij} = 1) + \phi \left\{ \sum_{j \in s_i} B_{ij}(I_{ij} = 1) \right\} \right)$$

where ϕ is the Huber function. Translating the idea for MQ we have:

$$\theta_i^{MQD} = N_i^{-1} \left(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} g^{-1}(\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\bar{q}_j}) - \sum_{h=1}^{m} \sum_{j \in s_h} \hat{B}_{jh}(I_{jh} = 1) + \phi \left\{ \sum_{h=1}^{m} \sum_{j \in s_h} \hat{B}_{jh}(I_{jh} = 1) \right\} \right). \tag{7}$$

The ϕ-function in MQD depends on a tuning constant c. Using min-max method to compute the optimal tuning constant we obtain

$$\theta_i^{MQD} = N_i^{-1} \left(\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} g^{-1}(\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\bar{q}_j}) - \frac{1}{2} \left(min \left\{ B_{jh}(I_{jh} = 1) \right\} + max \left\{ B_{jh}(I_{jh} = 1) \right\} \right) \right) \tag{8}$$

where the conditional bias for unit j has to be computed depending of the type of the response variable. If y_{ij} is a continuous

$$\hat{B}_{hj}(I_{hj} = 1) = \sum_{i \notin s_i} \mathbf{x}_{ij}^T \left\{ \sum_{h=1}^{m} \sum_{j \in s_h} \mathbf{x}_{hj} \hat{d}_{hj} \mathbf{x}_{hj}^T \right\}^{-1} \hat{d}_{hj} \mathbf{x}_{hj} (y_{hj} - \mathbf{x}_{hj}^T \hat{\boldsymbol{\beta}}_{\tau_i}). \tag{9}$$

3 Model-based simulations

In this section, we provide results regarding model-based simulation scenarios for continuous variables. We use a simulation setup based on Chambers et al. (2014). We consider the following outcome model for generating the finite population for $m = 40$ small areas:

$$y_{ij} = 100 + 5x_{ij} + u_i + \epsilon_{ij},$$

where i refers to the areas and j to the population units. Values for x are generated as i.i.d. from a lognormal distribution with a mean of 1 and a standard deviation of 0.5 on the log scale. The area and individual random effects are independently generated according to the following scenarios:

a) [0,0,0] - no outliers, $u \sim N(0, 3)$ and $e \sim N(0, 6)$;

b) [e,u,0] - outliers in area (fixed) and individual effects, $u \sim N(0,3)$ for areas 1–36, $u \sim N(9,20)$ for areas 37–40 and $e \sim \delta N(0,6) + (1-\delta)N(20,150)$.

The sample data are selected by a simple random sampling without replacement within each area. The population and sample size are the same for all areas and are fixed at $N_i = 100$ and $n_i = 5$.

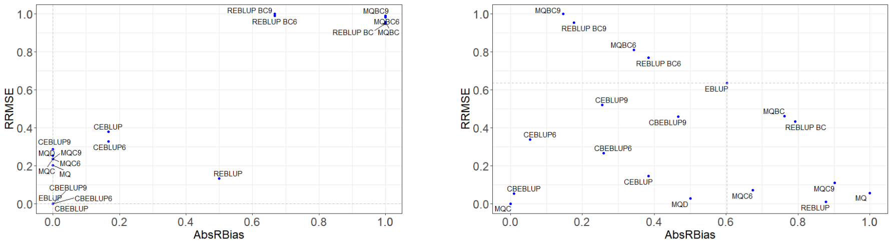
Each scenario is independently simulated 1000 times. The parameter of interest is the population mean in each small area. Nine different estimators are used for this purpose: the M-quantile estimator MQ by (Chambers and Tzavidis, 2006) which serves as a reference for the MQ regression based estimators, the bias corrected M-quantile estimator MQBC by (Chambers et al., 2014), the M-quantile estimator based on full bias correction MQC (see equation (4)), the M-quantile estimator based on conditional bias correction MQD (see equation (8)),the standard EBLUP which serves as a reference for all the considered estimators, the robust eblup REBPLUP by (Sinha and Rao, 2009) and its robust bias corrected version REBLUP–BC by (Chambers et al., 2014), the CBEBLUP and CEBLUP predictorS by (Dongmo-Jiongo et al., 2013). The influence function ϕ that is used in MQBC, MQC, REBLUP BC, CBEBLUP and CEBLUP is a Huber proposal 2 type. For each estimator, we test three different tuning constant for the bias correction part equal to 3, 6 and 9. The performance of the proposed indicators is evaluated according to min-max plots (Figure 1). The values on the x-axis and $y-$axis on plots are:

$$AbsRBias = \frac{\text{Median}[\text{AbsB}(\theta_{ki})] - \min\{\text{Median}[\text{AbsB}(\Theta_i)]\}}{\max\{\text{Median}[\text{AbsB}(\Theta_i)]\} - \min\{\text{Median}[\text{AbsB}(\Theta_i)]\}}$$

and

$$RRMSE = \frac{\text{Median}[\text{RRMSE}(\theta_{ki})] - \min\{\text{Median}[\text{RRMSE}(\Theta_i)]\}}{\max\{\text{Median}[\text{RRMSE}(\Theta_i)]\} - \min\{\text{Median}[\text{RRMSE}(\Theta_i)]\}},$$

where θ_{ki} is the kth estimator in the ith area and Θ_i is the vector all K predictors in area i.



(a) (0,0,0)

(b) (e,u,0)

Figure 1: Min-Max plots for MQ, MQBC, MQC, MQD, EBLUP, REBLUP, REBLUP BC, CBEBLUP and CEBLUP under selected simulation scenarios.

Results confirm our expectations regarding the behaviour of the MQC and MQD estimators. With respect to MQ estimator, the new proposed estimators reduce the bias in the presence of outliers.

References

Beaumont, J., Haziza, D., and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100:555–569.

Chambers, R., Chandra, H., Salvati, N., and Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B*, 76 (1):47–69.

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93 (2):255–268.

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81(396):1063–1069.

Dongmo-Jiongo, V., Haziza, D., and Duchesne, P. (2013). Controlling the bias of robust small area estimators. *Biometrika*, 100:843–858.

Sinha, S. K. and Rao, J. N. K. (2009). Robust small area estimation. *The Canadian Journal of Statistics*, 37 (3):381–399.

Best Prediction of Missing Area-Level Direct Estimates via Multivariate Modelling

Anna-Lena Wölwer [a*], Jan Pablo Burgard [b*], Domingo Morales [c**] and Ralf Münnich [d*]

[*]Department of Economic and Social Statistics, Trier University, Germany

[**]Operations Research Center, University Miguel Hernández de Elche, Spain

Abstract

In the last years one could see increasing methodological research and applications of multivariate Fay-Herriot (MFH) models. The models allow for various structures of random effects and sampling variances and can further improve the quality of the model-based predictions. In applications to real data, however, MFH models can suffer from partially missing direct estimates of the variables of interest. This can frequently occur when considering direct estimates from different survey or different points in time as dependent variables. Burgard et al. (2021b, 2019) introduce a variant of the bivariate Fay-Herriot model which allows for partially missing direct estimates. They present parameter estimation (ML and REML), derive (empirical) best predictors and approximations to the corresponding MSE for the new model. We extend their work on bivariate models to arbitrary multivariate models and missing structures of the variables of interest, conduct simulation studies, and give an application to publicly available data from the *American Community Survey* (ACS).

Keywords— area-level models, multivariate models, small area estimation, missing values

1 The multivariate FH model

For fine regional and demographic domains, direct survey estimates can be associated with high variability due to small sample sizes. Model-based small

[a]woelwer@uni-trier.de

[b]burgardj@uni-trier.de

[c]d.morales@umh.es

[d]muennich@uni-trier.de

area estimation (SAE) techniques facilitate to increase the effective sample size of domain-level direct estimates by combining similar domains in a common model-based framework; a procedure which is referred to as borrowing strength. A comprehensive overview of SAE techniques is given in Rao and Molina (2015) and Morales et al. (2021). There are two main types of model-based small area estimation techniques, unit- and area-level models. In the small area context their variants are often referred to as Battese-Harter Fuller (BHF) and Fay-Herriot (FH) models respectively, following the works of Battese et al. (1988) and Fay and Herriot (1979). Even though area-level models do not directly use unit-level sampling information, but only aggregate domain statistics, there are a number of reasons for their use, several of them listed in Morales et al. (2021, Chapter 16). For non-linear statistics, the auxiliary information have to be available for the entire target population at the unit-level. This is often not given or leads to the fact that valuable but only aggregated auxiliary information cannot be used. Furthermore, researchers often do not get access to unit-level information on fine regional and demographic levels, but only aggregate statistics.

In the class of area-level small area models, multivariate Fay-Herriot (MFH) models have received more attention in recent years. With MFH models several variables of interest are modelled simultaneously, additionally profiting from the correlation structure between them. One can model one statistic over different points in time or several statistics from the same survey together. The possibility of using additional information from the same survey for SAE motivated Fay (1987) to propose a multivariate version of the FH model. He applied the model to estimate the median income of three-, four-, and five-person households in the *U.S. Current Population Survey* (CPS). The structure of the MFH model accounts for covariances of the sampling errors which is especially necessary when considering variables of interest from the same survey. Even when one is interested in one variable alone, the multivariate modeling can further increase the precision of each variable of interest when the variables are sufficiently correlated. Thus, MFH models facilitate to include further variables of interest as well as (estimated) auxiliary information for which sampling error covariances can be estimated. Further early work with MFH models is given in Datta et al. (1991) and Datta et al. (1999). To name a few applications of the MFH model, in the context of poverty estimation it is applied in Huang and Bell (2004) with further studies in Huang and Bell (2006), Morales et al. (2015), Porter et al. (2015), Benavent and Morales (2016), Arima et al. (2017), Ubaidillah et al. (2019), Benavent and Morales (2021), and Burgard et al. (2021a). We refer to Benavent and Morales (2016) for a general description of the MFH model and its parameter estimation.

2 Partially missing information

Especially when using data from different sources it can frequently occur that information which is needed for modeling is partially missing. Area-level information can be partially missing when the domains of interest are not incorporated in the sampling design (via stratification) and thus - by chance - domain-specific sample sizes can be zero such that no direct estimate can be computed. Furthermore, statistical agencies usually only publish aggregate statistics for which a statistic of the variation, e.g. the standard error or the coefficient of variation, does not exceed a certain threshold. Molina and Marhuenda (2015) recall that in official statistics the threshold for the coefficient of variation is usually set to 20%. In addition to that, for disclosure control statistical agencies set minimum cell counts for the publication of frequency tables, see Hundepool et al. (2010) for an overview. The previously mentioned reasons can lead to missing values both in the variables of interest and auxiliary data. Partially missing auxiliary data can be imputed. Then, however, the associated measurement errors should be considered in the FH model. The use of partially-imputed auxiliary information motivated Lohr and Ybarra (2002) to investigate an extension of the FH model to measurement errors, known as the *measurement error model* which is published in Ybarra and Lohr (2008). The model is extended in Burgard et al. (2021a) to bivariate FH models and a bivariate normal distribution of measurement errors.

Next to the auxiliary data, also the direct estimates of interest may be partially missing. Then, a multivariate - or a corresponding univariate - FH model can only be applied to the domains with full information. Using a model fit on domains with complete data, synthetic predictions can be calculated for domains with missing direct estimates, see e.g. Morales et al. (2021, Chapter 16).

Let there be D domains and $m > 1$ dependent variables. Then, we can partition the set of domains in subsets $\mathcal{D}_0 = \{1, \ldots, D_0\}$ and $\mathcal{D}_1 = \{D_0 + 1, \ldots, D\}$, where $D_0 < D$, such that if the vector of the m direct estimates $\boldsymbol{y}_d$ is completely observed $d \in \mathcal{D}_0$ and if at least one entry of $\boldsymbol{y}_d$ is not observed $d \in \mathcal{D}_1$. Parameters $\boldsymbol{\beta}$, i.e. the fixed effects of the model, and $\boldsymbol{\theta}$, i.e. the variance parameters of the model, can be estimated based on information from $\mathcal{D}_0$ via maximum likelihood (ML) or restricted maximum likelihood (REML). The synthetic predictor of the characteristic of interest $\boldsymbol{\mu}_d$ is given by

$$\hat{\boldsymbol{\mu}}_d^{syn} = \boldsymbol{X}_d \hat{\boldsymbol{\beta}}, \quad d = 1, \ldots, D, \tag{1}$$

where $\boldsymbol{X}_d$ is the matrix of auxiliary information for domain d. For domains with missing direct estimates the mean squared error of the synthetic predictor can be approximated by

$$MSE(\hat{\boldsymbol{\mu}}_d^{syn}) \approx \boldsymbol{X}_d (\boldsymbol{X}_0^\top \boldsymbol{V}_0^{-1} \boldsymbol{X}_0)^{-1} \boldsymbol{X}_d^\top + \boldsymbol{V}_{ud}, \quad \forall d \in \mathcal{D}_1, \tag{2}$$

where quantities $\boldsymbol{X}_0$ and $\boldsymbol{V}_0$ are defined solely based on data from $\mathcal{D}_0$, compare Morales et al. (2021, 441–442). The MSE can be estimated by plugging in $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$.

Applying the MFH model only to the domains with complete information, however, is unsatisfactory. The estimation of the parameters (apart from the correlation of the random effects) can be worse than with the corresponding univariate FH models. This occurs when there are only few domains for which no observation is missing and when the missing pattern is heterogeneous across domains. On the other hand, the univariate FH model ignores the correlation of the variables of interest, thereby only using part of the available information. Furthermore, the synthetic predictor is not considering the information of other domain-specific direct estimates in a domain which could give valuable information for the prediction of the missing values.

3 The multivariate FH model under partially missing information

Burgard et al. (2021b, 2019) introduce a bivariate Fay-Herriot model under partially missing direct estimates of the dependent variables, called missing data BFH (MBFH) model. They give ML and REML fitting algorithms to estimate model parameters. Furthermore, they introduce empirical best predictors of target values and derive approximations to the mean squared error. For the bivariate case Burgard et al. (2021b, 2019) allow some of the direct estimates y_{dk}, $k = 1, 2$, to be missing. By setting $y_{\bar{d}1} = (y_{d1}, 0)^\top$ and $y_{\bar{d}2} = (0, y_{d2})^\top$, three groups of domains can be distinguished:

$\mathbb{D}_1 = \{d \in \mathbb{N} : 1 \leq d \leq D_1\}$ containing the D_1 domains where only y_{d1} is observed.

$\mathbb{D}_2 = \{d \in \mathbb{N} : D_1 + 1 \leq d \leq D_1 + D_2\}$ containing the D_2 domains where only y_{d2} is observed.

$\mathbb{D}_3 = \{d \in \mathbb{N} : D_1 + D_2 + 1 \leq d \leq D\}$ containing the remaining domains with fully observed $y_d = (y_{d1}, y_{d2})'$.

The best predictor (BP) of $\boldsymbol{u}_d$ under the MFH model, exemplary shown for domains in $\mathbb{D}_1$, is given by

$$\hat{\boldsymbol{u}}_d^{bp} = E[\boldsymbol{u}_d|\boldsymbol{y}_d] = \boldsymbol{\Phi}_{d1} \begin{pmatrix} \sigma_{ed1}^{-2} & 0 \\ 0 & 0 \end{pmatrix} (y_{\bar{d}1} - \boldsymbol{X}_d\boldsymbol{\beta}), d \in \mathbb{D}_1 \quad (3)$$

with

$$\boldsymbol{\Phi}_{d1} = \left[\begin{pmatrix} \sigma_{ed1}^{-2} & 0 \\ 0 & 0 \end{pmatrix} + \boldsymbol{V}_{ud}^{-1} \right]^{-1}. \quad (4)$$

By considering the partially-missing direct estimates in (3), domain-specific best predictions of random effects can be given, also for the missing direct estimates. This is a significant advantage of the MBFH model compared to the synthetic predictions which could else-wise only be calculated for missing values in a FH or MFH model.

We extend the model introduced in Burgard et al. (2021b, 2019) to multivariate dependent variables, derive empirical best predictors of domain parameters and approximations to the mean squared error. The derived algorithms are examined in model-based Monte Carlo simulation studies under different correlation settings of random effects and sampling errors. We furthermore apply the model to publicly available data from the *American Community Survey* (ACS) estimating the county-year median income of Hispanic or Latino Americans. The simulation studies and application reveal the flexibility and applicability of the proposed approach to different small area estimation problems.

References

Arima, S., Bell, W. R., Datta, G. S., Franco, C., and Liseo, B. (2017). Multivariate Fay-Herriot Bayesian estimation of small area means under functional measurement error. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4):1191–1209.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Benavent, R. and Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics & Data Analysis*, 94:372–390.

Benavent, R. and Morales, D. (2021). Small area estimation under a temporal bivariate area-level linear mixed model with independent time effects. *Statistical Methods & Applications*, 30(1):195–222.

Burgard, J. P., Esteban, M. D., Morales, D., and Pérez, A. (2021a). Small area estimation under a measurement error bivariate Fay-Herriot model. *Statistical Methods & Applications*, 30(1):79–108.

Burgard, J. P., Morales, D., and Wölwer, A.-L. (2019). Area-level small area estimation with missing values. Research Papers in Economics 2019-14, University of Trier, Department of Economics.

Burgard, J. P., Morales, D., and Wölwer, A.-L. (2021b). Small area estimation of socioeconomic indicators for sampled and unsampled domains. *AStA Advances in Statistical Analysis.* DOI:10.1007/s10182-021-00426-4.

Datta, G. S., Day, B., and Basawa, I. (1999). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference*, 75(2):269–279.

Datta, G. S., Fay, R. E., and Ghosh, M. (1991). Hierarchical and empirical multivariate Bayes analysis in small area estimation. In *Proceedings of the U.S. Bureau of the Census 1991 Annual Research Conference*, pages 63–79.

Fay, R. E. (1987). Application of multivariate regression to small domain estimation. In Platek, R., Rao, J. N. K., Särndal, C. E., and Singh, M. P., editors, *Small Area Statistics*, pages 91–102, New York. Wiley.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366):269–277.

Huang, E. T. and Bell, W. R. (2004). An empirical study on using ACS supplementary survey data in SAIPE state poverty models. In *2004 Proceedings of the American Statistical Association*, pages 3677–3684. US Bureau of the Census Washington DC.

Huang, E. T. and Bell, W. R. (2006). Using the t-distribution in small area estimation: an application to saipe state poverty models. In *Proceedings of the Survey Research Methods Section*, pages 3142–3149.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordholt, E. S., Seri, G., and de Wolf, P.-P. (2010). *Handbook on Statistical Disclosure Control*. ESSnet on Statistical Disclosure Control, version 1.2 edition. https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf.

Lohr, S. L. and Ybarra, L. M. R. (2002). Area-level models using data from multiple surveys. In *Proceedings of Statistics Canada Symposium*.

Molina, I. and Marhuenda, Y. (2015). sae: An R package for small area estimation. *The R Journal*, 7(1):81–98.

Morales, D., Esteban, M. D., Pérez, A., and Hobza, T. (2021). *A Course on Small Area Estimation and Mixed Models: Methods, Theory and Applications in R*. Springer.

Morales, D., Pagliarella, M. C., and Salvatore, R. (2015). Small area estimation of poverty indicators under partitioned area-level time models. *SORT: Statistics and Operations Research Transactions*, 39(1):19–34.

Porter, A. T., Wikle, C. K., and Holan, S. H. (2015). Small area estimation via multivariate Fay-Herriot models with latent spatial dependence. *Australian & New Zealand Journal of Statistics*, 57(1):15–29.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*, volume 2. John Wiley & Sons, Hoboken, New York.

Ubaidillah, A., Notodiputro, K. A., Kurnia, A., and Mangku, I. W. (2019). Multivariate Fay-Herriot models for small area estimation with application to household consumption per capita expenditure in Indonesia. *Journal of Applied Statistics*, 46(15):2845–2861.

Ybarra, L. M. R. and Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4):919–931.

Index of authors