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

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 $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{iK})'$ 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 $Y_i = \{Y_{i1}, \ldots, Y_{iN_i}\}$ 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 $p_{ij} = (p_{ij1}, \ldots, p_{ijK})$; these latter are modeled according to the following multinomial logit specification:

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

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 $\bar{Y}_i = (\bar{Y}_{i1}, \ldots, \bar{Y}_{iK})'$, 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 ($i = 1, \ldots, m, j \in s_i$), while covariates x_{ij} are available at the population level ($i = 1, \ldots, m, j = 1, \ldots, N_i$). 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 $p_i = (p_{i1}, \ldots, p_{ik})'$, 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 $\xi_g = (\xi_{g1}, \ldots, \xi_{gK})'$ with masses

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

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 $Y_{ij}, i = 1,\ldots,m, j =$

$1,\ldots,N_i$, 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.