STATISTICS IN TRANSITION
new series

An International Journal of the Polish Statistical Association

Survey Methodology

STATISTICS IN TRANSITION new series and ***SURVEY METHODOLOGY***

Small Area Estimation, Poznań 2014

Joint Issue Part 2

CONTENTS

Volume 17, Number 1, March 2016

Editorial Office **ISSN 1234-7655**
Marek Cierpiał-Wolan, Ph.D., Scientific Secretary
m.wolan@stat.gov.pl
Secretary:
Beata Witek, b.witek@stat.gov.pl
Agata Bara, a.bara@stat.gov.pl
Phone number 00 48 22 — 608 33 66
Rajmund Litkowiec, Technical Assistant
Address for correspondence
GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1 pp. 1–2

FROM THE EDITORS

This issue is devoted entirely to selected papers presented at the international conference on Small Area Estimation − SAE 2014. The conference took place at the Economics University of Poznan, from the 3rd to the 5th of September 2014. A satellite event − a workshop devoted to small area estimation with R given by Professor Li-Chun Zhang from the University of Southampton and Statistics Norway − was organized on the eve of the conference (September 2, 2014). The main aim of the SAE 2014 Conference was to provide a forum for the current research on small area estimation and related fields. The conference focused on aspects of conceptual, methodological and practical achievements in small area estimation methods in recent years. The conference brought together specialists from universities working on small area estimation, practitioners working in National Statistical Offices and other research agencies, all over the world. The SAE 2014 Conference was organized as part of activities of the European Working Group on Small Area Estimation and was the next in the series of conferences which have so far been held in Jyväskylä, Pisa, Elche and Trier.

The Programme Committee of the Conference was chaired by Professor Domingo Morales, Universidad Miguel Hernández de Elche. The Organizing Committee of the SAE 2014 Conference was chaired by Professor Marcin Szymkowiak, Poznan University of Economics. More details about these committees are available at this link: http://www.sae2014.ue.poznan.pl/index.html.

Statistics in Transition has previously published seven issues that focused on SAE, starting with articles from the Warsaw International Conference held in 1992 (*Vol. 1, Number 6, 1994*), and ending in 2005-6 with two issues (*Vol. 7, No. 3, December 2005*, and *Vol. 7, No. 4, March 2006*) with selected articles from The Conference held at the University of Jyväskylä, Finland, from 27-31 August 2005.

This time, in view of the large number of papers (15), the SAE 2014 proceedings were split into two parts: the first one appeared in December 2015. This is the second part of this thematic issue. These proceedings mark a turning as they were co-edited by Professor Włodzimierz Okrasa, Editor of Statistics in Transition, and Dr. Michael Hidiroglou, Editor of Survey Methodology. This joint editorship is a first between our journals, and it was a pleasure and memorable experience for both editors to collaborate.

These two issues represent a subset of the invited articles presented at the conference. They all went through a formal review process that was shared by four Guest Editors: Professor Risto Lehtonen (University of Helsinki), Finland,

Professor Ray Chambers (University of Wollongong), Australia, Dr. Graham Kalton (Westat), U.S.A., and Professor Malay Ghosh (University of Florida), U.S.A. Both editors, Professor Okrasa and Dr. Hidiroglou, are very grateful for the excellent collaboration and efficient work of the Guest Editors. Our appreciation goes also to authors, especially those who had directly collaborated with us or with our editorial offices on adjusting their papers to our journals' technical requirements.

It is with great satisfaction that we, as editors, provide the reader with such a unique collection of papers representing not only the state-of-the-art variety of small area estimation topics, but also a great deal of thoughtful suggestion for exploration in further research.

Michael Hidiroglou

Wlodzimierz Okrasa

Editors

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 3–6

FROM THE GUEST EDITORS (PART 2)

The second part of this Joint Issue of Statistics in Transition and Survey Methodology includes seven articles. These two issues have been split according to which guest editors have been looking after the articles. They are not necessarily sequenced according to the themes that appeared in the original Conference programme.

The first paper, by Erciulescu and Fuller, presents a small area procedure where the mean and variance of an auxiliary variable are subject to estimation error. They consider fixed and random specifications for these auxiliary variables. Their study was motivated by a situation where the sample used for small area estimation was a subsample of a larger survey. The larger survey furnished estimates of the distribution of the auxiliary variables. They demonstrate that efficiency gains associated with the random specification for the auxiliary variable measured with an error can be obtained. They propose a parametric bootstrap procedure for the mean squared error of the predictor based on a logit model. The resulting bootstrap procedure has a smaller bootstrap error than a classical double bootstrap procedure with the same number of samples.

The second paper, by Münnich, Burgard, Gabler, Ganninger and Kolb, develops a sampling design that can support accurate estimation for the 2011 German Census. In contrast to carrying out a classical census, a register-assisted census, using population register data and an additional sample, was implemented. The main objective of the census was to produce the total population counts at fairly low levels of geography. Ralf Münnich et al. provide an overview of how the sampling design recommendations were set up to fulfill legal requirements and to guarantee an optimal, yet flexible, source of information. Small area methods, as well as traditional methods, were used to produce these counts. Empirical results of the small area estimation are presented.

The next three papers present developments in small area estimation methodology and practical application in various fields of empirical research and statistics production, including poverty research and fisheries statistics. The first paper, by Guadarrama, Molina and J. N. K. Rao, provides a review on methods for the estimation of poverty indicators for small areas, including design-based direct estimation and a number of model-based small area estimation methods: the Fay-Herriot area level model, the World Bank poverty mapping method (the ELL method) and three Bayesian variants previously published by the authors. These are the empirical best/Bayes (EB) and hierarchical Bayes (HB) methods and a Census EB method providing an extension of the EB method. While the

Fay-Herriot method employs area-level data, the other methods require unit-level auxiliary information. The ELL, EB, Census EB and HB methods rely on statistical data infrastructures where access to unit-level records of population units taken for example from administrative registers and population censuses is available for research and statistics production. This option is becoming frequently met in an increasing number of countries and much of current small area research is conducted under this assumption. The list of advantages and disadvantages, reported for each of the methods, appears helpful for practitioners facing the challenge of choosing a small area method for a particular estimation task. Statistical properties (bias and accuracy) of methods are assessed empirically by model-based simulation experiments with unit-level synthetic data following a nested error model, throwing further light on the methodological summaries of the methods. Extensive simulation scenarios of varying complexity include informative sampling and a nested error model with outliers; these scenarios in particular are important for practical purposes. For practical application, it is important that also situations are considered where some of the underlying assumptions of the methods do not hold, which is often the case in practice. The conclusions drawn by the authors on the relative performance of the methods are useful for researchers and practitioners.

Because of its applicability in various data infrastructures, the Fay-Herriot model has been widely used in small area estimation purposes all over the world and new developments are often needed to extend the method for practical situations at hand. A robust hierarchical Bayesian approach for the Fay-Herriot area-level model is presented in the second paper, written by Chakraborty, Datta and Mandal. The starting point is the authors' observation on a possible poor performance of the standard Fay-Herriot area-level model in the presence of outliers. The new method is aimed for cases where extreme values are met for some of the random effects of small area means, causing problems in the standard Fay-Herriot procedure under normality assumptions of the random effects. The authors propose a two-component normal mixture model, which is based on noninformative priors on the model variance parameters, regression coefficients and the mixing probability. The method is aimed as an alternative to a scale mixture of normal distributions with known mixing distribution for the random effects. The authors apply their method to real data of US Census Bureau for poverty rate estimation at county level. The results indicate that probabilities of having large random effects are expected to be low for most areas but can be large for some areas, thus calling for attention to handle the possible heterogeneity of the data. Simulation studies based on artificially generated data are conducted to assess the performance of the proposed method against the standard Fay-Herriot model. In the first set of experiments, the authors verify the robustness of the proposed method to outliers in the cases considered. In further simulations, the authors show that their method tends to perform better than the Fay-Herriot method when the possibility of presence of outliers is high, and performs similarly in situations where outliers are not expected. In their concluding notes

the authors provide a useful discussion on the possible causes of exceptionally large random effects for certain areas, calling for a careful specification of the linking model and the choice of the explanatory (auxiliary) variables.

The third paper, by Hernandez-Stumpfhauser, Breidt and Opsomer, provides a refinement of the Fay-Herriot approach for a particular small area estimation problem. The authors consider a practical problem of developing a new weighting procedure for a regular fisheries survey in the United States on recreational fishing in saltwater. For the estimation of the recreational catch, fishing catch per trip is estimated from one survey and the number of fishing trips from another survey. Data from these two surveys are combined to estimate recreational fishing catch in 17 US states. For weighting procedure, estimates are needed for the fraction of fishermen who leave the fishing site during a prespecified time interval on a selected day. The distribution of daily departure times is needed within spatio-temporal domains subdivided by mode of fishing. Direct estimates could be obtained but they are not sufficient because of a large number of estimation domains, causing very small (even zero) domain sample sizes. The authors develop a small area estimation solution based on the Fay-Herriot approach. More specifically, the authors show that with a certain hierarchical model formulation that is slightly more complex as the standard mixed model, fast and accurate model selection procedure based on variational/Laplace approximation to the posterior distribution can be implemented for the particular estimation problem considered. Even if the underlying linear mixed model can be complex involving fixed and random effects for the states, waves and fishing modes and interaction terms, the method can serve as a cost-effective alternative to the computationally more demanding MCMC sampler. By empirical comparison of MCMC and the proposed variational/Laplace approaches using real data, the authors show that the results are essentially identical, thus motivating the use of the method in practice.

The production of small area statistics by national statistical agencies and international statistical institutes is becoming more and more important for societal planning and evaluation and the allocation of public funds to regional areas and other population subgroups. In the next paper, Kordos presents a personal view on the development of certain aspects of small area estimation methodology and practice in the context of official statistics. The author first summarizes the main approaches in small area estimation with some historical remarks. He continues by discussing the important issue of the use of administrative records in official statistics production and as auxiliary information in the construction of estimators for various regional indicators. The author presents a summary of international conferences on small area estimation organized in past years, covering a period from 1985. Further, he presents a review of selected international small area estimation programs and research projects on small area estimation. A special property of these research activities is that they are conducted in cooperation with research communities on small area estimation and actors whose responsibility is in the production of official small

area statistics. The interaction has proven fruitful in motivating ongoing research and development in small area estimation methodology and for boosting the implementation of methods in regular official statistics production. This aspect might well be taken as the main message of the paper by Kordos.

In many national statistical institutes, the design-based approach has offered the prevailing paradigm in official statistics production for decades. Good reasons are the ability of the approach to provide estimates having favorable statistical properties such as design-unbiasedness, which is often appreciated by the clients, and the availability of powerful statistical procedures and tools that use effectively the auxiliary information supplied in various forms. Calibration techniques and generalized regression estimation are examples of such methods. While relative standard errors of design-based estimates can be sufficiently small for population domains whose sample size is large, this is not necessarily the case for small domains. It is in this field of action where model-based small area estimation is challenging the design-based approach. In the final paper, Hidiroglou and Estevao present an empirical assessment of selected design-based methods against some existing model-based small area estimation methods, considered at Statistic Canada. Traditional design-based estimators include the Horvitz-Thompson estimator, two variants of calibration estimators and a modified regression estimator. A synthetic estimator and the standard EBLUP and its variant called pseudo-EBLUP represent model-based methods. The relative performance of the methods is assessed in design-based simulation experiments, where in addition to "ideal" conditions also misspecified models are considered. The relative performance of the methods differs depending on whether the model holds or not. Of the traditional design-based estimators, the domain-specific calibration estimator and the modified regression estimator indicate the best efficiency. The model-based small area estimators tend to outperform the design-based methods in efficiency, especially for small domains. As expected, the model-based methods can suffer from large design bias in cases where the model is misspecified.

Several persons (in addition to the Editor and Guest Editors) have served as reviewers of papers published in this thematic issue of the journal. We acknowledge the efforts of F. Jay Breidt, Isabel Molina, Domingo Morales, Ari Veijanen, Mamadou Diallo and Jon Rao: their encouraging and productive comments directly contributed to the quality of the papers.

Risto Lehtonen and **Graham Kalton**

Guest Editors

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 7−8

SUBMISSION INFORMATION FOR AUTHORS

Statistics in Transition new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and the Central Statistical Office of Poland, on a quarterly basis (during 1993–2006 it was issued twice and since 2006 three times a year). Also, it has extended its scope of interest beyond its originally primary focus on statistical issues pertinent to transition from centrally planned to a market-oriented economy through embracing questions related to systemic transformations of and within the national statistical systems, world-wide.

The SiT-*ns* seeks contributors that address the full range of problems involved in data production, data dissemination and utilization, providing international community of statisticians and users – including researchers, teachers, policy makers and the general public – with a platform for exchange of ideas and for sharing best practices in all areas of the development of statistics.

Accordingly, articles dealing with any topics of statistics and its advancement – as either a scientific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series.*

Demonstration of the role played by statistical research and data in economic growth and social progress (both locally and globally), including better-informed decisions and greater participation of citizens, are of particular interest.

Each paper submitted by prospective authors are peer reviewed by internationally recognized experts, who are guided in their decisions about the publication by criteria of originality and overall quality, including its content and form, and of potential interest to readers (esp. professionals).

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl.,
GUS / Central Statistical Office
Al. Niepodległości 208, R. 287, 00-925 Warsaw, Poland

It is assumed, that the submitted manuscript has not been published previously and that it is not under review elsewhere. It should include an abstract (of not more than 1600 characters, including spaces). Inquiries concerning the submitted manuscript, its current status etc., should be directed to the Editor by email, address above, or w.okrasa@stat.gov.pl.

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: http://stat.gov.pl/en/sit-en/guidelines-for-authors/

Survey Methodology is an internationally acclaimed scientific journal that is published twice a year. For over 40 years, it has been a source of key information on survey methods for statisticians. Survey Methodology draws on the expertise of statisticians and experts from Canada and around the world. It provides reliable, complete and authoritative information.

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Survey Methodology is published twice a year in electronic format. Submitted articles are peer reviewed by experts in the particular area that the author(s) address.

Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor:

statcan.smj-rte.statcan@canada.ca,
Statistics Canada, 150 Tunney's Pasture Driveway,
Ottawa, Ontario, Canada, K1A 0T6

For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/SurveyMethodology).

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 9–24

SMALL AREA PREDICTION UNDER ALTERNATIVE MODEL SPECIFICATIONS

Andreea L. Erciulescu[1], **Wayne A. Fuller**[2]

ABSTRACT

Construction of small area predictors and estimation of the prediction mean squared error, given different types of auxiliary information are illustrated for a unit level model. Of interest are situations where the mean and variance of an auxiliary variable are subject to estimation error. Fixed and random specifications for the auxiliary variables are considered. The efficiency gains associated with the random specification for the auxiliary variable measured with error are demonstrated. A parametric bootstrap procedure is proposed for the mean squared error of the predictor based on a logit model. The proposed bootstrap procedure has smaller bootstrap error than a classical double bootstrap procedure with the same number of samples.

Key words: unit level model, parametric bootstrap, double bootstrap, measurement error, auxiliary information.

1. Introduction

Small area estimation procedures use models and auxiliary data to construct estimates for subpopulations that are more efficient than the direct estimators for those subpopulations. Modeling provides potential for gains by postulating a distribution for the unknown parameters. The presence of variables that are correlated with the variable of interest provides potential for efficiency gains when there is knowledge about the distribution of those variables. In most of the small area literature the small area population means of the auxiliary variables are assumed to be known. We are interested in the situation where only estimates of the parameters of the distribution of the auxiliary variables are available. Our study was motivated by a situation where the sample used for small area estimation was a subsample of a larger survey. The larger survey furnished estimates of the distribution of the auxiliary variables.

[1] National Institute of Statistical Sciences and USDA NASS, 1400 Independence Ave. SW, Room 6040 F, Washington, DC 20250.
[2] Iowa State University, 1214 Department of Statistics, Ames, IA 50010.

A number of papers consider measurement error in the auxiliary variables used in the linear regression model. See Fuller and Harter (1987), Ghosh, Sinha and Kim (2006), Ghosh and Sinha (2008), Torabi, Datta and Rao (2009), Ybarra and Lohr (2008) and Datta, Rao and Torabi (2010). In contrast, we study unit level mixed models where the observed explanatory variables are measured without error, but the parameters of the distribution of the auxiliary variables are known subject to estimation error. We consider auxiliary information obtained from a sample, including the limit case of a complete sample.

Because there are no closed-form estimators for the prediction mean squared error (MSE) for most nonlinear models, bootstrap methods have been suggested. See Hall and Maiti (2006) and Pfeffermann and Correa (2012). We propose parametric bootstrap procedures based on the work of Davidson and MacKinnon (2007).

This paper is organized in sections. In Section 2.2 we present predictors of small area means assuming a unit level generalized linear mixed model, with alternative specifications for the auxiliary information. In Section 2.4 we describe parametric double bootstrap procedures for MSE estimation. Section 3.2 contains simulation results comparing the prediction MSEs for the logit model under alternative model specifications and alternative types of data for the auxiliary variables. Simulation comparisons of alternative bootstrap prediction MSE estimators are given in Section 3.3.

2. Unit Level Nonlinear Models

2.1. Introduction

The unit level generalized linear mixed model considered in this study is

$$E\left[y_{ij}|\boldsymbol{x}_{ij}, b_i\right] = g(\boldsymbol{x}_{ij}\boldsymbol{\beta}, b_i), \tag{1}$$

$$\boldsymbol{x}_{ij} = \boldsymbol{\mu}_{xi} + \boldsymbol{\varepsilon}_{ij}, \tag{2}$$

$i = 1, \ldots, m$, where m is the number of areas, j is the index for units in the area, $\boldsymbol{\beta}$ is a vector of coefficients, $\boldsymbol{\mu}_{xi}$ is the area mean of the auxiliary variable, and b_i is the area random effect. It is assumed that the b_i are independent and identically distributed, with a density f_b with mean 0 and variance σ_b^2, mutually independent of ε_{ij}, where the ε_{ij} are independent and identically distributed random variables with a density f_ε with mean 0 and variance σ_ε^2. The vector $(y_{ij}, \boldsymbol{x}_{ij}), i = 1, \ldots, m, j = 1, \ldots, n_i$ is observed.

Additional information on the distribution of $\boldsymbol{x}_{ij}$ may be available. Possibilities include a second sample of $\boldsymbol{x}_{ij}$ observations, or an estimator of $\boldsymbol{\mu}_{xi}$,

or complete knowledge of the distribution function. The area means of $\boldsymbol{x}$ can be treated as fixed or as random variables. If random, we assume

$$\boldsymbol{\mu}_{xi} = \boldsymbol{\mu}_{x} + \boldsymbol{\delta}_{i}, \tag{3}$$

where δ_i are independent and identically distributed, with a density f_δ with mean 0 and variance σ_δ^2. Assume δ_i are independent of $b_k, e_{ij}, \varepsilon_{rt}$, for all i, k, r and t, where $e_{ij} = y_{ij} - g(\boldsymbol{x}_{ij}\boldsymbol{\beta}, b_i)$.

Of interest is the i^{th} small area mean of $\boldsymbol{y}$

$$\theta_i = \int g(\boldsymbol{x}\boldsymbol{\beta}, b_i) dF_{\boldsymbol{x}_i}(\boldsymbol{x}), \tag{4}$$

where $F_{\boldsymbol{x}_i}(\boldsymbol{x})$ is the distribution of $\boldsymbol{x}$ in area i. Also of interest is the prediction mean squared error

$$\alpha_i = E(\hat{\theta}_i - \theta_i)^2, \tag{5}$$

where $\hat{\theta}_i$ is the predictor. We assume throughout that the area population is large so that we need not consider finite population corrections.

The nature of the estimation-prediction problem is determined by the distributional properties of the vector $(b_i, \boldsymbol{\delta}_i, \boldsymbol{\varepsilon}_{ij})$. The nonlinear model is more complicated than the linear model for several reasons. First, parameter estimation is more difficult because no closed form estimator exists. Likewise, closed form estimators of the mean squared error do not exist. Lastly, the small area mean of the auxiliary variable is not sufficient for the estimation of θ_i.

As an example of model (1), consider a Bernoulli response variable $\boldsymbol{y}$, with realizations y_{ij} for m different areas and n_i different units within each area. To simplify the presentation, we consider scalar x_{ij} for the remainder of our discussion. Let x_{ij} be independent and identically distributed, following a distribution $F_{\boldsymbol{x}_i}$. Let the expected value of y_{ij} given $(\boldsymbol{x}_{ij}, b_i)$ be

$$g(\boldsymbol{x}_{ij}\boldsymbol{\beta}, b_i) = \frac{exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i)}{1 + exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i)}, \tag{6}$$

where $\boldsymbol{x}_{ij} = (1, x_{ij})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. The model is the generalized linear mixed model with logit link.

2.2. Predictors of $\boldsymbol{\theta_i}$

We present predictors of θ_i for model (6), under alternative specifications for $\boldsymbol{x_{ij}}$ and for different levels of auxiliary information, given known parameters $(\sigma_b^2, \sigma_\varepsilon^2, \sigma_\delta^2, \boldsymbol{\beta}, \mu_x)$.

2.2.1. Known Covariate Distribution

Let the distribution of x_{ij} be known and let $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ be a random sample of (x_{ij}, y_{ij}), where $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n_i}), \boldsymbol{y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,n_i})$. Then, given known parameters, the minimum mean squared error (MMSE) predictor of the i^{th} small area mean of $\boldsymbol{y}$ is

$$\begin{aligned} \hat{\theta}_i &= E[\theta_i(b)|(\boldsymbol{x}_i, \boldsymbol{y}_i)] \\ &= \frac{\int_b \theta_i(b) \prod_{t=1}^{n_i} f(y_{it}|x_{it}, b) dF_b(b)}{\int_b \prod_{t=1}^{n_i} f(y_{it}|x_{it}, b) dF_b(b)}, \end{aligned} \tag{7}$$

where

$$\theta_i(b) = \int g(\boldsymbol{x}\boldsymbol{\beta}, b) dF_{\boldsymbol{x}_i}(\boldsymbol{x}).$$

In some finite population situations, the entire finite population of $\boldsymbol{x}$ values may be known and the integral expression for $\theta_i(b)$ in (7) is the sum over the population. In the simulations for this model we assume $x_{ij} \sim NI(\mu_{xi}, \sigma_\varepsilon^2)$ with μ_{xi} known and σ_ε^2 known.

2.2.2. Sample Estimated Covariate Distribution

Let an estimator of the distribution of x_{ij} be given by a sample $(x_{ij}, w_{ij}), j = 1, \ldots, r_i$, where w_{ij} are weights such that the sample cumulative distribution function (CDF) is unbiased for the population CDF. Then, given known $(\sigma_b^2, \boldsymbol{\beta})$, the predictor of the i^{th} small area mean of $\boldsymbol{y}$ is

$$\begin{aligned} \hat{\theta}_i &= E[\theta_i(b)|(\boldsymbol{x}_i, \boldsymbol{y}_i)] \\ &= \frac{\int_b \theta_i(b) \prod_{t=1}^{n_i} f(y_{it}|x_{it}, b) dF_b(b)}{\int_b \prod_{t=1}^{n_i} f(y_{it}|x_{it}, b) dF_b(b)}, \end{aligned} \tag{8}$$

where

$$\theta_i(b) = \sum_{j=1}^{r_i} w_{ij} g(\boldsymbol{x_{ij}}\boldsymbol{\beta}, b).$$

The sample used to estimate the CDF could be the original sample with $r_i = n_i$ or the estimation sample could be the original sample augmented by an

additional probability sample of size n_i' selected from the area population. See Ghosh et al. (2009) for an example using the sample CDF.

2.2.3. Unknown Random Covariate Mean

Assume the form of the distribution of $\boldsymbol{x}$ for area i is known, with unknown parameters $(\mu_{xi}, \sigma_\varepsilon^2)$. Assume μ_{xi} satisfies (3). Then, given known $(\sigma_b^2, \sigma_\varepsilon^2, \sigma_\delta^2, \boldsymbol{\beta}, \mu_x)$, the MMSE predictor of the i^{th} small area mean of $\boldsymbol{y}$ is

$$\begin{aligned} \hat{\theta}_i &= E[\theta_i(b,\delta)|(\boldsymbol{x}_i, \boldsymbol{y}_i)] \\ &= \frac{\int_b \int_\delta \theta_i(b,\delta) \prod_{t=1}^{n_i} f(y_{it}|x_{it},b) f(x_{it}|\delta) dF_\delta(\delta) dF_b(b)}{\int_b \int_\delta \prod_{t=1}^{n_i} f(y_{it}|x_{it},b) f(x_{it}|\delta) dF_\delta(\delta) dF_b(b)}, \end{aligned} \tag{9}$$

where

$$\theta_i(b,\delta) = \int g[(\mu_x + \delta + \boldsymbol{\varepsilon})\,\boldsymbol{\beta}, b] dF_{\boldsymbol{\varepsilon}_i}(\boldsymbol{\varepsilon}).$$

In the simulations we assume $x_{ij} \sim NI(\mu_{xi}, \sigma_\varepsilon^2)$ and $\delta_i \sim NI(0, \sigma_\delta^2)$.

2.2.4. Unknown Random Covariate Mean, Additional Information $\tilde{\boldsymbol{x}}_i$

Let the random model assumptions of Section 2.2.3 hold. Let a vector of n_i' observations on x_{ij}, denoted by $\tilde{\boldsymbol{x}}_i$, be available. Then, given known $(\sigma_b^2, \sigma_\varepsilon^2, \sigma_\delta^2, \boldsymbol{\beta}, \mu_x)$, the MMSE predictor of the i^{th} small area mean of $\boldsymbol{y}$ is

$$\begin{aligned} \hat{\theta}_i &= E[\theta_i(b,\delta))|(\boldsymbol{x}_i, \boldsymbol{y}_i, \tilde{\boldsymbol{x}}_i)], \\ &= \frac{\int_b \int_\delta \theta_i(b,\delta) \prod_{t=1}^{n_i} f(y_{it}|x_{it},b) f(x_{it}|\delta) \prod_{t'=1}^{n_i'} f(\tilde{x}_{it'}|\delta) dF_\delta(\delta) dF_b(b)}{\int_b \int_\delta \prod_{t=1}^{n_i} f(y_{it}|x_{it},b) f(x_{it}|\delta) \prod_{t'=1}^{n_i'} f(\tilde{x}_{it'}|\delta) dF_\delta(\delta) dF_b(b)}, \end{aligned}$$

where

$$\theta_i(b,\delta) = \int g[(\mu_x + \delta + \boldsymbol{\varepsilon})\boldsymbol{\beta}, b]\, dF_{\boldsymbol{\varepsilon}_i}(\boldsymbol{\varepsilon}).$$

In the simulations we assume $\tilde{x}_{ij'} \sim NI(\mu_{xi}, \sigma_\varepsilon^2)$, so $\tilde{\mu}_{xi} = (n_i')^{-1} \sum_{j'=1}^{n_i'} \tilde{x}_{ij'}$ is a sufficient statistic for μ_{xi} and the predictor simplifies to

$$\hat{\theta}_i = \frac{\int_b \int_\delta \theta_i(b,\delta) \prod_{t=1}^{n_i} f(y_{it}|x_{it},b) f(x_{it}|\delta) f(\tilde{\mu}_{xi}|\delta) dF_\delta(\delta) dF_b(b)}{\int_b \int_\delta \prod_{t=1}^{n_i} f(y_{it}|x_{it},b) f(x_{it}|\delta) f(\tilde{\mu}_{xi}|\delta) dF_\delta(\delta) dF_b(b)}. \tag{10}$$

2.3. Estimation

In practice, the vector of parameters $\boldsymbol{\psi} = (\sigma_b^2, \sigma_\varepsilon^2, \sigma_\delta^2, \boldsymbol{\beta}, \mu_x)$ is not known and needs to be estimated. Consider the model specified by (1), (2), (3), (6), with additional information $\tilde{\boldsymbol{x}}_i$ available, as described in Section 2.2.4. The likelihood is

$$L(\sigma_b^2, \sigma_\varepsilon^2, \sigma_\delta^2, \boldsymbol{\beta}, \mu_x | \boldsymbol{x}, \boldsymbol{y}, \tilde{\boldsymbol{x}}) = \prod_{i=1}^{m} L_i,$$

where

$$\begin{aligned} L_i &= \int_b \int_\delta \prod_{j=1,j'=1}^{n_i,n_i'} f(y_{ij}, x_{ij}, \tilde{x}_{ij'} | b, \delta, \psi) f(b|\psi) f(\delta|\psi) d\delta db \\ &= \int_b \prod_{j=1}^{n_i} f(y_{ij}|b, x_{ij}, \beta) f(b|\sigma_b^2) db \int_\delta \prod_{j=1}^{n_i+n_i'} f(x_{ij}^*|\delta, \mu_x, \sigma_\varepsilon^2) f(\delta|\sigma_\delta^2) d\delta, \end{aligned}$$

and $\boldsymbol{x}^* = (\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is the vector of all available auxiliary information.

Notice that the likelihood $L(\sigma_b^2, \sigma_\varepsilon^2, \sigma_\delta^2, \boldsymbol{\beta}, \mu_x | \boldsymbol{x}, \boldsymbol{y}, \tilde{\boldsymbol{x}})$ factors into $L(\sigma_b^2, \boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{x})$ and $L(\sigma_\varepsilon^2, \sigma_\delta^2, \mu_x | \boldsymbol{x}, \tilde{\boldsymbol{x}})$. Hence, the parameters $(\sigma_\varepsilon^2, \sigma_\delta^2, \mu_x)$ can be estimated separately from the estimation of the parameters $(\sigma_b^2, \boldsymbol{\beta})$. Estimation of $(\sigma_\varepsilon^2, \sigma_\delta^2, \mu_x)$ can be based on maximizing the likelihood for the linear mixed model specified in (2) and (3), with additional information $\tilde{\boldsymbol{x}}_i$ available.

Numerical integration methods are required for construction of estimates and predictions.

2.4. Bootstrap MSE Estimation

In this section we consider estimation of the MSE of $\hat{\theta}_i$ as a predictor of θ_i. Let $\boldsymbol{\psi}$ be the parameter that defines the distribution of the sample observations, and let $\widehat{\boldsymbol{\psi}}$ be an estimator of $\boldsymbol{\psi}$. Let $\boldsymbol{\alpha}$ be a vector of parameters of interest and let $\boldsymbol{\alpha}^*$ be a parametric bootstrap (simulation) estimator of $\boldsymbol{\alpha}$. For the models considered in Section 2.2, let α_i be the MSE of the prediction error for area i, as defined in (5). For the nonlinear small area model with known distribution for x_{ij}, the vector of parameters is $\boldsymbol{\psi} = (\sigma_b^2, \boldsymbol{\beta})$. For the nonlinear small area models with unknown random μ_{xi}, the vector of parameters is $\boldsymbol{\psi} = (\sigma_b^2, \boldsymbol{\beta}, \sigma_\varepsilon^2, \mu_x, \sigma_\delta^2)$. Because there is no closed form expression for the prediction MSE given in (5), we consider bootstrap MSE estimation.

A sample generated with $\boldsymbol{\psi}$ and random number seed r is said to be created with data generator $(\boldsymbol{\psi}, r)$, denoted $DG(\boldsymbol{\psi}, r)$. Let B_1 bootstrap samples be generated using random number seeds $r_{1,1}, r_{1,2}, \ldots, r_{1,B_1}$. Let $\boldsymbol{\psi}_k^*$ be the estimator of $\boldsymbol{\psi}$ from the kth bootstrap sample generated using $DG(\widehat{\boldsymbol{\psi}}, r_{1,k})$. The bootstrap estimator of prediction MSE for area i is

$$\hat{\alpha}_i^* = B_1^{-1} \sum_{k=1}^{B_1} (\hat{\theta}_{i,k}^* - \theta_{i,k}^*)^2 =: B_1^{-1} \sum_{k=1}^{B_1} \alpha_{i,k}^* = \bar{\alpha}_i^*, \qquad (11)$$

where $\theta_{i,k}^*$ is the true small area mean generated for the kth bootstrap sample, $\hat{\theta}_{i,k}^*$ is the sample predictor of $\theta_{i,k}^*$ and $\alpha_{i,k}^*$ is the prediction squared error for the kth bootstrap sample. The estimator (11) is called the level-one bootstrap estimator.

In the double bootstrap, a sample estimator, denoted by α_i^{**}, is generated using $\boldsymbol{\psi}^*$ from the level-one generated sample. Typically a large number of α_i^{**} is generated for each α_i^* and the bias adjusted estimator is

$$\tilde{\alpha}_i^{**} = B_1^{-1} \sum_{k=1}^{B_1} 2\alpha_{i,k}^* - B_1^{-1} B_2^{-1} \sum_{k=1}^{B_1} \sum_{t=1}^{B_2} \alpha_{i,k,t}^{**}. \tag{12}$$

where $\alpha_{i,k,t}^{**}$ is generated using $DG(\boldsymbol{\psi}_k^*, r_{2,k,t})$, B_1 is the number of level-one bootstrap samples, B_2 is the number of level-two bootstrap samples per level-one sample, and the $r_{2,k,t}, k = 1,2,\ldots,B_1, t = 1,2,\ldots,B_2$, are independent random numbers, independent of $r_{1,k}$.

We use a double bootstrap estimator based on the work of Davidson and MacKinnon (2007) who give a fast double bootstrap procedure for bootstrap testing. See also Giacomini, Politis and White (2013). In the fast double bootstrap, a single α_i^{**} is generated for each α_i^*. Let $r_{2,1}, r_{2,2}, \ldots, r_{2,B_1}$ be a second independent sequence of random numbers. Given the sequence of random numbers, define $\alpha_{i,k}^{**}$ to be calculated from data generated with $DG(\boldsymbol{\psi}_k^*, r_{2,k})$. The (classic) double bootstrap estimator used in this study is

$$\tilde{\alpha}_{i,C}^{**} = B_1^{-1} \sum_{k=1}^{B_1} (2\alpha_{i,k}^* - \alpha_{i,k}^{**}) = 2\bar{\alpha}_i^* - \bar{\alpha}_i^{**}. \tag{13}$$

To construct an even more efficient bootstrap estimator, define $\alpha_{i,k,2}^*$ to be calculated from data generated with $DG(\widehat{\boldsymbol{\psi}}, r_{2,k})$. Then a bias adjusted (double bootstrap) estimator is

$$\hat{\alpha}_i^{**} = B_1^{-1} \sum_{k=1}^{B_1} (\alpha_{i,k}^* + \alpha_{i,k,2}^* - \alpha_{i,k}^{**}), \tag{14}$$

where the quantity $\alpha_{i,k}^{**} - \alpha_{i,k}^*$ is a one-degree-of-freedom estimator of the bias. If one uses $r_{2,1}$ as $r_{1,2}$, $r_{2,2}$ as $r_{1,3}$, etc., a form of (14) becomes

$$\tilde{\alpha}_{i,T}^{**} = B_1^{-1} \sum_{k=1}^{B_1} (\alpha_{i,k}^* + \alpha_{i,k+1}^* - \alpha_{i,k}^{**}), \tag{15}$$

where $\alpha_{i,k+1}^*$ is generated with $DG(\widehat{\boldsymbol{\psi}}, r_{1,k+1})$ and $\alpha_{i,k}^{**}$ is generated with $DG(\boldsymbol{\psi}_k^*, r_{1,k+1})$. We call the estimator (15) a telescoping bootstrap because it is of the form (14) using lagged values of $\alpha_{i,k}^*$. If the use of $r_{2,k}$ in place of an

independent random number results in positive correlation between $\alpha_{i,k}^{*}$ and $\alpha_{i,k-1}^{**}$, then $\tilde{\alpha}_{i,T}^{**}$ will have smaller simulation variance than $\tilde{\alpha}_{i,C}^{**}$ of (13).

3. Simulations

In the simulation study we consider $m = 36$ areas with unit level observations x_{ij} in three groups of 12 areas, with sizes $n_i \in \{2,10,40\}$. The number of additional unit level observations is $n_{i,} = 10$, for each area i. Each sample, $(\boldsymbol{y}, \boldsymbol{x}, \tilde{\boldsymbol{x}})$, is generated using model (1 - 3) with $\sigma_b^2 = 0.25, \mu_x = 0, \sigma_\delta^2 = 0.16$, and $\sigma_\varepsilon^2 = 0.36$. The vector of coefficients for the fixed effects is $(\beta_0, \beta_1) = (-0.8,1)$ and $\boldsymbol{x}_{ij} = (1, x_{ij})$. For each unit, the probability that $y_{ij} = 1$ is

$$g(\boldsymbol{x_{ij}\beta}, b_i) = \frac{\exp(-0.8+x_{ij}+b_i)}{1+\exp(-0.8+x_{ij}+b_i)}. \quad (16)$$

The population mean of $g(\boldsymbol{x_{ij}\beta}, b_i)$ is 0.334 with variance 0.029. An area with $\mu_{xi} = 0.4$ has mean 0.412 with variance 0.028. Four hundred Monte Carlo samples were generated satisfying the model.

The estimation models are:

- Model 1: Specified by (1) and (6) and described in Section 2.2.1. Known normal distribution for x_{ij}. The distribution of y_{ij} is

$$f(y_{ij}|x_{ij}, b_i) = I(y_{ij}, 1)g(\boldsymbol{x_{ij}\beta}, b_i) + I(y_{ij}, 0)(1 - g(\boldsymbol{x_{ij}\beta}, b_i)),$$

  where $I(y_{ij}, .)$ is the indicator function, and $g(\boldsymbol{x_{ij}\beta}, b_i)$ is defined in (16). The distribution of b_i is $N(0,0.25)$.

- Model 2: Specified by (1) and (6) and described in Section 2.2.2. Sample estimated distribution of $\boldsymbol{x}$ based on the original sample $\boldsymbol{x}$.
- Model 2*: Specified by (1) and (6) and described in Section 2.2.2. Sample estimated distribution of $\boldsymbol{x}$ based on the original sample $\boldsymbol{x}$ augmented by a sample $\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_m)$.
- Model 3: Specified by (1), (2), (6) and described in Section 2.2.3. Unknown random auxiliary mean μ_{xi}. Distributions of y_{ij} and b_i are the same as those for Model 1. The distribution of x_{ij} is defined by the random model given in Section 2.2.3.
- Model 4: Specified by (1), (2), (3), (6) and described in Section 2.2.4. Unknown random auxiliary mean μ_{xi} and observed $\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_m)$.

The models are fitted as generalized linear mixed models, using the *lmer* and *glmer* functions in the *lme4* package in R. The true i^{th} small area mean of $\boldsymbol{y}$ is given by (4) and the predicted i^{th} area means of $\boldsymbol{y}$ are given in (7 - 10), with estimated $(\mu_x, \beta_0, \beta_1, \sigma_b^2, \sigma_\delta^2, \sigma_\varepsilon^2)$. The integrals in (4, 7 - 10) were approximated using a 26-point approximation to the normal distribution.

3.1. Refinement of Prediction MSE Estimators

Wang and Fuller (2003) suggested the estimator of σ_δ^2 be bounded by

$$K_{\delta,s} = 0.5\left[\hat{V}(\hat{\sigma}_\delta^2 | \sigma_\delta^2 = 0)\right]^{0.5},$$

where $\hat{V}(\hat{\sigma}_\delta^2 | \sigma_\delta^2 = 0)$ is the estimated variance of $\hat{\sigma}_\delta^2$, given $\sigma_\delta^2 = 0$. Because of the large degrees of freedom for $\hat{\sigma}_\varepsilon^2$, we set $K_{\delta,s}$ equal to the true value of 0.008 in the simulations,

$$K_{\delta,s} = 0.5[2m(m-1)^{-1}(\textstyle\sum_{i=1}^{m}((n_i + n_{i\prime})^{-1}\sigma_\varepsilon^2)^{-2})^{-1}]^{0.5} = 0.008.$$

Similarly, we bound the estimator of σ_b^2 by

$$K_{b,s} = 0.5\left[V(\hat{\sigma}_b^2 | \sigma_b^2 = 0)\right]^{0.5} = 0.006.$$

The proportion of sample estimators $\hat{\sigma}_b^2$ that hit the bound is 0.025, the proportion of level one estimators of $\hat{\sigma}_b^{2*}$ that hit the bound is 0.111. If $\hat{\sigma}_{b,k}^2 = 0.006$ we set $\alpha_{i,k}^{**}$ equal to $\alpha_{i,k}^{*}$. That is, the estimated bias is zero for such samples.

Using (13), one can obtain an unacceptable double bootstrap prediction MSE estimator, where the estimated bias for a sample is greater than the estimate. In practice, one would increase the number of bootstrap samples. Rather than build such a procedure into our Monte Carlo algorithm, we defined bounds for the estimator. Thus, the final estimator is

$$\hat{\alpha}_{i,C}^{**} = \begin{cases} 1.60\bar{\alpha}_i^{*}, & if \quad \bar{\alpha}_i^{*-1}\bar{\alpha}_i^{**} > 1.60 \\ 0.83\bar{\alpha}_i^{*}, & if \quad \bar{\alpha}_i^{*-1}\bar{\alpha}_i^{**} < 0.83 \\ \tilde{\alpha}_{i,C}^{**}, & otherwise \ , \end{cases} \tag{17}$$

where 0.83 and 1.60 are the 0.025 and 0.975 points of the chi-square distribution with 199 $(B_1 - 1)$ degrees of freedom, and $\tilde{\alpha}_{i,C}^{**}$ is defined in (13). The analogous definition holds for the telescoping estimator of (14). See Hall and Maiti (2006) for an alternative definition of the direct double bootstrap estimates.

The proportions of sample estimators of $\hat{\alpha}_{i,T}^{**}$ that hit the lower bound defined in (17) are 0.016, 0.016 and 0.013, for the areas of sizes 2, 10 and 40, respectively. The proportions of sample estimators of $\hat{\alpha}_{i,T}^{**}$ that hit the upper bound defined in (17) are 0.026, 0.069 and 0.084, for the areas of sizes 2, 10 and 40, respectively. Due to larger variability in the classic double bootstrap estimators, the proportions of sample estimators of $\hat{\alpha}_{i,C}^{**}$ that hit the lower bound defined in (17) are 0.058, 0.048 and 0.041, for the areas of sizes 2, 10 and 40, respectively, and the proportions of sample estimators of $\hat{\alpha}_{i,C}^{**}$ that hit the upper bound defined in (17) are 0.155, 0.201 and 0.183, for the areas of sizes 2, 10 and 40, respectively.

3.2. MSE for Different Types of Auxiliary Information

The coefficient of variation for $\hat{\sigma}_b^2$ calculated for the 400 Monte Carlo samples is about 0.64, approximately the CV of a Chi-square with five degrees of freedom. The Monte Carlo relative bias of the estimator of $\hat{\sigma}_b^2$ is about -0.12, which is approximately equal to eighteen Monte Carlo standard errors.

Table 1 contains estimates of the prediction MSE, denoted by α, for fixed and random models with different amounts of auxiliary information. The simulation MSE standard errors are presented in parantheses below the MSE values. The smallest MSE is for Model 1, where the covariate distribution is known. The next smallest MSE is for Model 4, where the form of the covariate distribution is known, the covariate mean is random and the auxiliary information is available. The largest MSE is for Model 2, where the covariate distribution is not specified. The small area mean predictor for Model 3 is the conditional expected value formula given in (9). Notice that in the construction of the small area predictor for Model 4, given in (10), the conditioning is also on the additional source of information, $\boldsymbol{x}$, available for the areas.

The extra observations on x_{ij} represent additional information available about the distribution of $\boldsymbol{x}$ for the area. Hence, the large gain in efficiency associated with $\boldsymbol{x}$ for sample size two (compare 10.94 for Model 2^* to 17.29 for Model 2). Model 3 differs from Model 2 in that the distribution of x_{ij} is assumed to be normal and the area mean is also assumed to be normally distributed. Adding these distributional assumptions changes the MSE from 17.29 to 13.22 for sample size two. The effect of added information is smaller for the random μ_{xi} models (models 2^* and 4) than for the fixed μ_{xi} models (models 2 and 3).

The contribution of the variance of the estimation error in the mean of $\boldsymbol{x}$ to the MSE depends on the importance of $\boldsymbol{x}$ in the model and on the size of the samples. With $n_i = 2$, the MSE with known area mean of $\boldsymbol{x}$ is 57% of the MSE with no additional information on the distribution of $\boldsymbol{x}$. The reduction in MSE from adding independent observations on $\boldsymbol{x}$ is related to the sizes of the two samples and to the model. If the small area mean of $\boldsymbol{x}$ is fixed, the original sample is ten observations and the added sample is ten observations, the MSE falls midway

between that with no additional information and that with complete information. With fixed small area mean of $\boldsymbol{x}$, an original sample of size 2 and an added sample of size 10, the expected added variance is one sixth of that of the original sample. In this simulation the effect of treating the mean as random is equivalent to adding 2.25 observations on $\boldsymbol{x}$.

Table 1. MSE for different types auxiliary information (entries multiplied by 10^3)

Size	$\bar{y}$	Model 1	Model 2	Model 2*	Model 3	Model 4
2	102.14 (6.13)	9.88 (0.71)	17.29 (1.24)	10.94 (0.79)	13.22 (0.92)	10.72 (0.76)
10	20.15 (1.40)	7.15 (0.52)	8.56 (0.63)	7.87 (0.57)	8.26 (0.60)	7.76 (0.56)
40	5.14 (0.37)	3.46 (0.25)	3.81 (0.27)	3.74 (0.27)	3.78 (0.27)	3.72 (0.27)

Model 1: known distribution for x_{ij},
Model 2: unknown distribution for x_{ij}, with no $\tilde{\boldsymbol{x}}$,
Model 2*: unknown distribution for x_{ij}, with observed $\tilde{\boldsymbol{x}}$,
Model 3: random μ_{xi}, with no $\tilde{\boldsymbol{x}}$,
Model 4: random μ_{xi}, with observed $\tilde{\boldsymbol{x}}$

3.3. Monte Carlo Properties of Prediction MSE Estimators

The relative performances of bootstrap prediction MSE estimators under the different types of auxiliary information are similar. Therefore, we only present properties of prediction MSE estimators for Model 4, where the area mean μ_{xi} is random and auxiliary information $\tilde{\boldsymbol{x}}$ is available.

Table 2 contains results for $(\hat{\alpha}^{*}, \hat{\alpha}_T^{**}, \hat{\alpha}_C^{**})$ for the three area sample sizes, in groups of five lines. Each line is the average of the results for the 12 areas with the same sample size. The first line is the Monte Carlo estimates of the prediction MSE, $\hat{\alpha}$. The next four lines are of the bias relative to the mean, the coefficient of variation, the bias relative to the standard deviation and the bias relative to the standard error. The definitions are

$$RelBias = \sum_{is=1}^{12} \left(\hat{\alpha}_{.,is}^{EST} - \hat{\alpha}_{.,is}\right) \Big/ \sum_{is=1}^{12} \hat{\alpha}_{.,is} \,,$$

$$CV = \sum_{is=1}^{12} \sqrt{(400-1)^{-1} \sum_{\zeta=1}^{400} (\hat{\alpha}_{\zeta,is}^{EST} - \hat{\alpha}_{.,is}^{EST})^2} \Big/ \sum_{is=1}^{12} \hat{\alpha}_{.,is} \,,$$

$$\frac{Bias}{sd} = \sum_{is=1}^{12} \left(\hat{\alpha}_{.,is}^{EST} - \hat{\alpha}_{.,is}\right) \Bigg/ \sum_{is=1}^{12} \sqrt{(400-1)^{-1} \sum_{\zeta=1}^{400} (\hat{\alpha}_{\zeta,is}^{EST} - \hat{\alpha}_{.,is}^{EST})^2} \ ,$$

$$\frac{Bias}{se} = Bias/(20sd) \ ,$$

where ζ indexes the Monte Carlo samples, i denotes an area from a group of areas of sample size s, $\hat{\alpha}_{.,is} = (400)^{-1} \sum_{\zeta=1}^{400} \hat{\alpha}_{\zeta,is}$ is the average of the Monte Carlo prediction error estimators, $\hat{\alpha}_{.,is}^{EST} = (400)^{-1} \sum_{\zeta=1}^{400} \hat{\alpha}_{\zeta,is}^{EST}$ is the average of the bootstrap prediction MSE estimators, and $\hat{\alpha}^{EST} \in \{\hat{\alpha}^{*}, \hat{\alpha}_T^{**}, \hat{\alpha}_C^{**}\}$ is the bootstrap estimator for an area. The estimated prediction MSEs have CVs of about $40\%, 32\%$ and 22% for 200 bootstrap samples for sample sizes 2, 10, and 40, respectively.

In all cases the telescoping double bootstrap, denoted with a subscript T, has lower MSE than the classic double bootstrap, denoted with a subscript C. The estimators $\hat{\alpha}_T^{**}$ and $\hat{\alpha}_C^{**}$ have the same bias if the bound (17) is not used. The double bootstrap reduces the absolute value of the bias for all the sample sizes. However, the absolute bias of the double bootstrap is about 6% of the true value for sample size 2.

Table 2. Monte Carlo properties of prediction MSE estimators
($B_1 = 200, B_2 = 1$ and 400 MC samples, variances multiplied by 10^3)

Size	*Measure*	$\hat{\alpha}^{*}$	$\hat{\alpha}_T^{**}$	$\hat{\alpha}_C^{**}$
2	$V(\hat{\theta} - \theta)$	10.723	10.723	10.723
	RelBias	-0.143	-0.058	-0.062
	$CV(\hat{\alpha})$	0.403	0.456	0.477
	Bias/sd	-0.355	-0.127	-0.130
	Bias/se	-7.097	-2.537	-2.609
10	$V(\hat{\theta} - \theta)$	7.758	7.758	7.758
	RelBias	-0.133	-0.032	-0.039
	$CV(\hat{\alpha})$	0.318	0.365	0.385
	Bias/sd	-0.417	-0.087	-0.102
	Bias/se	-8.336	-1.738	-2.034
40	$V(\hat{\theta} - \theta)$	3.721	3.721	3.721
	RelBias	-0.082	0.016	0.009
	$CV(\hat{\alpha})$	0.222	0.260	0.286
	Bias/sd	-0.372	0.062	0.032
	Bias/se	-7.430	1.249	0.636

The variance of an estimator of the prediction MSE has two components. The first, that we call *between*, is the variance one would obtain if one used an infinite number of bootstrap samples. The second, that we call *within*, is the variability due to the fact that our set of bootstrap samples is a sample of samples.

We estimate these two components using two independent sets of bootstrap samples. That is, for each Monte Carlo sample, we generate two sets of $(B_1 = 100, B_2 = 1)$ samples. The sequences of random seeds $r_{1,k}', r_{2,k}', k = 1, \ldots, B_1$ for the second set are independent of the sequences of random seeds $r_{1,k}, r_{2,k}, k = 1, \ldots, B_1$ for the first set. Let $(\hat{\alpha}^*, \hat{\alpha}^{**}, \hat{\alpha}_T^{**}, \hat{\alpha}_C^{**})$ be the prediction MSE estimates for the first group of bootstrap samples and let $(\hat{\alpha}_2^*, \hat{\alpha}_2^{**}, \hat{\alpha}_{T2}^{**}, \hat{\alpha}_{C2}^{**})$ be the prediction MSE estimates for the second group of bootstrap samples. The within variance component for $B_1 = 100$ is estimated by half of the mean of squared differences between the two prediction MSE estimates,

$$Var_{within}^{EST} = (12)^{-1} \sum_{is=1}^{12} \left((400)^{-1} \sum_{\zeta=1}^{400} (\hat{\alpha}_{\zeta,is}^{EST} - \hat{\alpha}_{2,\zeta,is}^{EST})^2 \right)/2.$$

The variance components for the prediction MSE estimators $(\hat{\alpha}^*, \hat{\alpha}_T^{**}, \hat{\alpha}_C^{**})$ are given in Table 3 for $(B_1 = 100, B_2 = 1)$. The estimated between variance component is the difference between the estimated total variance and the estimated within variance component. The entries in the table are averages over the areas of the same sample size and over the Monte Carlo samples.

Table 3. Estimated variance components for variance of estimated prediction MSE
(Within is for 100 bootstrap samples. All variances have been multiplied by 10^6)

Source of Variation	Size	α^*	α_T^{**}	α_C^{**}
Between	2	17.886	23.040	23.040
Within		2.099	3.903	10.599
Total		19.985	26.943	33.639
Between	10	5.562	7.267	7.267
Within		1.099	2.324	5.376
Total		6.661	9.591	12.643
Between	40	0.544	0.725	0.725
Within		0.264	0.613	1.300
Total		0.808	1.338	2.025

The between component for the level one bootstrap is about 75% of the between component for the double bootstrap procedures. This is not surprising as bias reduction procedures often increase the variance. The bootstrap sampling variance, the within component, for the classic double bootstrap is about five times that of the level one bootstrap. The telescoping bootstrap is 2.1 to 2.7 times as efficient as the classic double bootstrap.

4. Summary

We used a simulation study of a unit level logistic model to compare the impact of different levels of auxiliary information. The minimum mean squared error predictors for the small area means were obtained by conditioning on the information available for an area. That information is the unit level response realizations, the unit level covariate observations, and the sometimes available additional unit level auxiliary information. We considered fixed and random mean models for the covariates, as well as known and unknown distribution for the covariates. The percentage effect on the prediction MSE of including auxiliary information in the estimation is smaller for the random mean model than for the fixed mean model for the covariates because using a random model is equivalent to adding observations.

We presented a parametric double bootstrap procedure for the prediction MSE for the unit level logistic model. The fast double bootstrap procedure, where the number of level-two bootstrap samples is $B_2 = 1$, has superior bootstrap efficiency relative to the classic double bootstrap procedure with $B_2 > 1$. The double bootstrap reduces the prediction MSE estimation bias to less than 50% of the bias of the level-one bootstrap. The double bootstrap increases the standard error of the prediction MSE estimator by 13 to 17% relative to that of the level-one bootstrap.

Acknowledgement

The work of Erciulescu was conducted while she was a student in the Departments of Statistics at Iowa State University. This research was partially supported by USDA NRCS CESU agreement 68-7482-11-534.

REFERENCES

DATTA, G. S., RAO, J. N. K., SMITH, D., (2005). On measuring the variability of small area estimators under a basic area level model, Biometrika, 92, 183−196.

DATTA, G. S., RAO, J. N. K., SMITH, D., (2012). Amendments and Corrections: On measuring the variability of small area estimators under a basic area level model, Biometrika, 99, 2, 509.

DATTA, G. S., RAO, J. N. K., TORABI, M., (2010). Pseudo-empirical Bayes estimation of small area means under a nested error linear regression model with functional measurements errors, Journal of Statistical Planning and Inference, 140, 2952−2962.

DAVIDSON, R., MACKINNON, J. G., (2007). Improving the reliability of bootstrap tests with the fast double bootstrap, Computational Statistics and Data Analysis, 51, 3259−3281.

FULLER, W. A., HARTER, R. M., (1987). The multivariate components of variance model for small area estimation, Small Area Statistics: An International Symposium, Platek, R., Rao, J. N. K., Sarndal, C.E. and Singh, M.P. (Eds.), John Wiley, New York, 103−123.

GHOSH, M., KIM, D., SINHA, K., MAITI, T., KATZOFF, M., PARSONS, V. L., (2009). Hierarchical and Empirical Bayes small domain estimation and proportion of persons without health insurance for minority subpopulations, Survey Methodology, 35, 53−66.

GHOSH, M., SINHA, K., (2007). Empirical Bayes estimation in finite population sampling under functional measurement error models, Journal of Statistical Planning and Inference, 137, 2759−2773.

HALL, P., MAITI, T., (2006). On parametric bootstrap methods for small area prediction, J.R. Statist. Soc. B, 68, 2, 221−238.

PFEFFERMANN, D., CORREA, S., (2012). Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation, Biometrika, 99, 457−472.

TORABI, M., DATTA, G., RAO, J. N. K., (2009). Empirical Bayes Estimation of Small Area Means under a Nested Error Linear Regression Model with Measurement Errors in the Covariates, Scandinavian Journal of Statistics, 36, 355−368.

WANG, J., FULLER, W. A., (2003). The mean squared error of small area spedictors constructed with estimated area variances, Journal of the American Statistical Association, 98, 716−723.

YBARRA, L. M. R., LOHR, S. L., (2008). Small area estimation when auxiliary information is measured with error, Biometrika, 95, 919−931.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 25–40

SMALL AREA ESTIMATION IN THE GERMAN CENSUS 2011

Ralf Münnich[1], Jan Pablo Burgard[2], Siegfried Gabler[3], Matthias Ganninger[4], Jan-Philipp Kolb[5]

ABSTRACT

In 2011, Germany conducted the first census after the reunification. In contrast to a classical census, a register-assisted census was implemented using population register data and an additional sample. This paper provides an overview of how the sampling design recommendations were set up in order to fulfil legal requirements and to guarantee an optimal but still flexible source of information. The aim was to develop a design that fosters an accurate estimation of the main objective of the census, the total population counts. Further, the design should also adequately support the application of small area estimation methods. Some empirical results are given to provide an assessment of selected methods. The research was conducted within the German Census Sampling and Estimation research project, financially supported by the German Federal Statistical Office.

Key words: register-assisted census, small area estimation, design optimisation, relative root mean squared error.

1. Introduction

The Census 2011 was the first after the German reunification. The last census in the Federal Republic of Germany was implemented in 1987, whereas the last census in the former German Democratic Republic (GDR) was conducted in 1981. For the first time in German Census history and for the first common census after the reunification, it was decided to conduct the Census in 2011 as a register-assisted census. The main sources of information are population registers. Additionally, a sample of approximately 10% of the population is drawn for two purposes. First, to assess the number of over- and under-counts in the registers aiming at deriving

[1] University of Trier. muennich@uni-trier.de.
[2] University of Trier. burgardj@uni-trier.de.
[3] GESIS Mannheim. siegfried.gabler@gesis.org.
[4] Roche Diagnostics. matthias@ganninger.de.
[5] GESIS Mannheim. jan-philipp.kolb@gesis.org.

the total census counts. Second, the sample information is used to estimate variables that were not included in the population registers. Certainly, the population register information can also be used as a source of auxiliary information.

The lowest level of official territorial division in Germany is the communities that have varying numbers of inhabitants. At the time of the Census 2011 the most populated community is Berlin but there are also 5 communities with less than 20 inhabitants. The most important target of the German Census was the determination of the official population sizes for each of the 11,399 communities. Due to the new census mode, adequate methodologies had to be developed, including sampling design and estimation strategies. Hence, a research project was granted by the Federal Ministry of the Interior and the German Federal Statistical Office to investigate an appropriate sampling design taking into account the German administration structure.

In addition to developing and recommending an optimal sampling design under the given circumstances, estimation strategies had to be developed that can be used in connection with this sampling design. In order to appropriately investigate the interplay of sampling design and estimation strategies, a close-to-reality universe of synthetic data had to be developed which was based on real register data. This universe was used as a sound basis for carrying out an extensive simulation. This article addresses the key findings of the sampling and estimation research project.

2. Objectives and frame of the German Census 2011

The Census 2011 sample had to be drawn to fulfil two main objectives:

Objective 1 Determination of the official population size for each community, i.e. estimating census over- and under-counts in order to derive the population sizes,

Objective 2 Estimation of key figures for additional variables.

Extensive planning preceded the realisation of the Census 2011. A census test, implemented in 2001, served as a preparation to gain initial information for the concept of a register-assisted census in Germany. The census law was launched in 2006. In the contract of the coalition, the reduction of burden and the use of modern methods were stipulated. The aim was to reduce costs without losing quality of important figures. The resulting figures should serve as a basis for administrative planning and decisions, and especially for financial adjustments between federal states. Therefore, it was necessary to reach a high level of quality.

The census law covered several important settings of the register-assisted census like variables of interest, rough description of the register-assisted structure, the sampling units, and quality margins. The quality constraints were especially important for objective 1 due to the importance of the population figures. The second objective was the estimation of variables not contained in the registers, e.g. on housing and living conditions. The relevant source of information for estimating over- and under-counts as well as for non-register variables is based on a sample of

addresses drawn from an address register containing all buildings and dwellings. Here, an address is defined as an address with housing space. There are addresses with only one inhabitant but also addresses with several hundred flats and inhabitants. A detailed description of the frame is provided by Kleber et al. (2009) and Bechtold (2013).

Numerous legal and administrative criteria of constraints had to be considered for the development of the underlying sampling design. As sampling units, complete addresses had to be drawn from the address register, i.e. all persons and households living at the given address. The address register was built exclusively for the Census 2011. In Germany, in general, one house is considered as an address. Obviously, the sampling units differed in size considerably, i.e. the variation of the number of inhabitants was very high which may have yielded a clustering effect for sampling. As an upper bound, 7.9 million inhabitants were sampled which covers approximately 10% of the population, not necessarily of the addresses. The design had to be as efficient as possible while considering the accuracy objectives for the estimation of the population counts stated in the census law. Further, feasibility was an important criterion that had to be considered.

Finally, the estimation had to be carried out for small areas and domains. The main areas of interest were districts or communities with at least 10,000 inhabitants. As domains, the main population subgroups were of interest. One of the main tasks at the beginning of the project was to find a coherent way of defining areas for sampling that considered the hierarchical structure of 16 federal states, 412 districts, and 11,339 communities.

As already mentioned, the sizes of communities in Germany differed greatly. Within the census law, it was stated that communities with at least 10,000 inhabitants played a major role in administrative and planning processes such that a different kind of inspection of over- and under-counts had to take place. It was important to consider these differences in the sampling design. Therefore, the first step was to build the so-called sampling points (SMP). These sampling points should be units with at least 10,000 registered persons that yield a frame of areas from which samples were drawn, according to the following scheme:

- Type 0 (SDT): Parts of communities with more than 400,000 inhabitants,
- Type 1 (GEM): Communities with at least 10,000 inhabitants and not of type 0,
- Type 2 (VBG): Collection of small communities within districts that together covered 10,000 inhabitants and more,
- Type 3 (KRS): Collection of the rest of small communities within a district.

With these settings, Germany was completely split into regional structures that considered all administrative and legal constraints and which could be used directly for optimizing the sampling design. The distribution of the sampling point types in Germany is illustrated by Figure 1. Sampling points of type 0 are depicted in white,

sampling points of type 1 are coloured dark grey. Sampling points of type 2 are coloured light grey and sampling points of type 3 are black in colour.

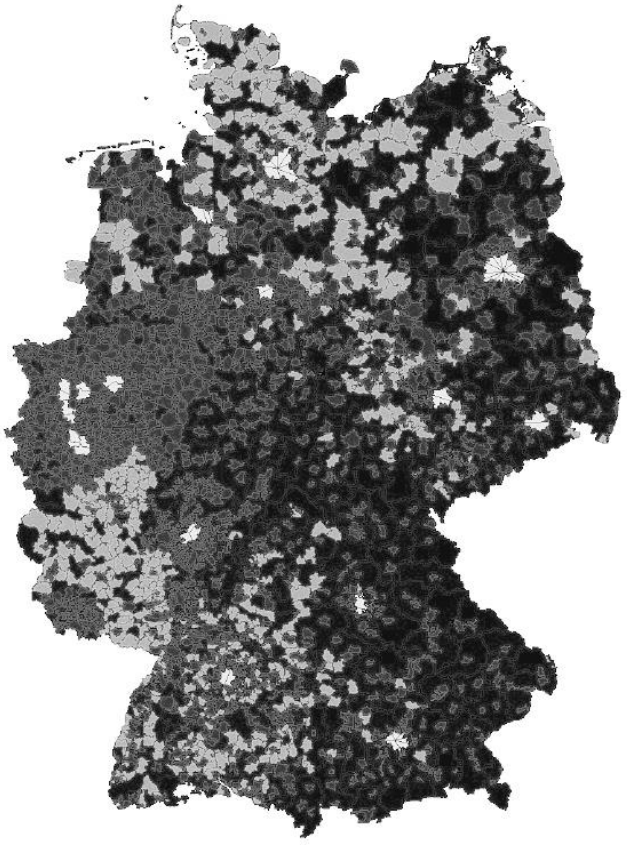


Figure 1. Map with sampling points in Germany (types 0 to 3)

3. Design optimisation and small area estimation

The main objective, stated before as objective 1, was to accurately determine the population counts. Nevertheless, it was necessary to keep regional and substantive points resulting from objective 2 in mind. The task, after all, was to derive an appropriate sampling design and to allocate the total sample size in an appropriate way to the aforementioned sampling points to fulfil certain quality specifications laid down by law. In order to appropriately account for quality margins in terms of relative variances or related components, design-based (or model-assisted) methods should be considered. However, with respect to local area analysis it was important to ensure that model-based, and particularly small area estimation methods, could be employed and not be adversely affected by a sampling design which is too elaborate.

3.1. Design optimisation

The criteria for the evaluation of the possible designs resulted from different sources. The requirements imposed by the German government were formalised in the census law. The target of the Federal States was to get reliable regional estimates and tables. The interest of academia was empirical research using Census microdata. There was a strong interest in model building based on results of the Census in economic and social sciences.

The selected survey design should ensure that different precision requirements were met for the different hierarchical entities. Adequate estimators had to be chosen to meet the requirements of the sampling design. The first step was to define the accuracy objectives adequately. This was necessary to decide on the allocation and optimisation issues of the sampling design.

As a starting point, the relative root mean squared error (RRMSE) was chosen as a measure of accuracy. It allows for the comparison of design-based and model-based estimators in a design-based environment. This collapses to the coefficient of variation for design-unbiased estimators. Based on the RRMSE, the following accuracy requirements were formulated for the Census 2011. The first objective concerned only total estimates $\hat{t}d$ of the size of the population U_d in communities d with more than 10,000 inhabitants:

$$\mathrm{RRMSE}(\hat{t}d) \leq 0,5\% \tag{1}$$

The same accuracy requirement was valid for parts of large towns with more than 400,000 inhabitants, the sampling points of type 0. As regards the second objective, the accuracy requirements depended on the type of sampling point and the variable of interest. In order to appropriately define the quality margins for objective 2 variables, the proportion p of the occurrence of an outcome of the variable of interest Y is used. The following rule was applied for all types of variables, whereas the proportion p varied across sampling point types. The proportion p of the variable of interest Y is given by

$$\frac{t_{dY}}{t_{dZ}} \approx p \quad \text{with} \quad p \geq \frac{1}{15} \tag{2}$$

where t_{dY} is the sum of inhabitants with property Y in area d and t_{dZ} is the total number of inhabitants in area d. Small proportions with $p < \frac{1}{15}$ were not considered under these settings. The accuracy requirement on the variable of interest is:

$$\mathrm{RRMSE}(\hat{t}_{dY}) \leq \frac{1}{p} \tag{3}$$

The relevant proportions p of the variables of interest are depicted in Table 1.

Table 1. Maximum RRMSE dependent on p

Objective	1	2	2	2	2	2	2
p (in %):	100	80	50	30	20	10	6.7
Maximum RRMSE (in %):	0.5	1.25	2	3.33	5	10	15

As already stated, an optimised sampling design had to satisfy all administrative criteria as well as the accuracy margins given above. However, some additional issues became important. Before 2011, the annual Microcensus sample of 1% was often used as a gold standard. Any census estimate of households and persons should not, therefore, be based on a smaller sampling fraction than that of the Microcensus. Further, it was necessary to ensure that mean squared error estimation should be possible in a closed form, at least for objective 1 estimates. Additionally, the design had to be robust against the above settings. Furthermore, considerable dissimilarities in the treatment of different groups of persons had to be reduced as much as possible. And finally, statistical modelling, like sociometric, econometric and, of course, small area models, should be supported.

In the context of model building, which is of particular interest for economic and social sciences, Gelman (2007) illustrated difficulties of survey weighting for regression modelling and argued that *survey weighting is a mess*. The *Gelman bound* (GB), which is defined as the ratio of the largest to the smallest design weight, is aimed not to exceed 10 and is unacceptable beyond 100. The reason for this is that Bayesian model building may become complicated in the presence of highly varying survey weights.

On the basis of the exigencies defined above, a stratified sampling design was suggested. Information on variances and the numbers of persons within addresses (objective 1) within the strata were available from the population register. Note that a comparison of the accuracy of different sampling designs is presented in Section 4.1, which yielded the recommendation to apply a stratified design.

Maximal sampling fractions had to be chosen because sample sizes within strata should not exceed the population sizes. Minimal sampling fractions should guarantee reliable estimates in all relevant areas. The approach published by Gabler et al. (2012) takes into account all of the above criteria. An optimal allocation in the Neyman-Tschuprov sense was developed, which satisfied the upper and lower bounds of the sample sizes within each stratum and, hence, is called box-constraint optimal allocation. This approach also allows the optimization of the sample sizes amongst all sampling points simultaneously using a 2-norm of the RRMSE for all areas of interest:

$$\| RRMSE(\hat{t}) \|_2 = \sqrt[2]{\sum_d \mathrm{RRMSE}(\hat{t}_d)^2}. \tag{4}$$

A comparison of different algorithms for the box-constraint optimal allocation can also be found in Münnich et al. (2012b) and, as an integer problem, in Friedrich et al. (2015).

In order to achieve a stratified sampling routine that enables a considerable variance reduction, all sampling points were stratified into eight address size classes. The eight classes in each sampling point were constructed to contain approximately the same number of persons. The box constraints yielded a maximal Gelman bound of 25.

3.2. Design-based and model-based small area estimation

Different estimators for the total of persons living in Germany have been examined within the research project. The most important ones are briefly presented here. An extensive discussion on these estimators is given in Rao (2003). Further details about the implementation in the German Census can be found in Münnich et al. (2012a). Münnich et al. (2009) discussed the application of binomial mixed-models and spatial small area models in the context of the census. For the implementation in other research projects see the working papers of the EURAREA project (see for example The EURAREA Consortium, 2004, or Guiblin et al., 2004) and the DACSEIS project (cf. Münnich et al., 2004).

The following estimators are considered in this paper:

- **Horvitz-Thompson estimator (HT)** The HT was considered as a benchmark. However, for objective 1, the loss of efficiency was very high since the population register was a very strong auxiliary variable.
- **Generalized regression estimator (GREG)** With regards to the GREG, the question arose of the level at which the parameter estimation for the regression coefficients should take place. Two major results appeared. First, a separation with regards to the address size class yielded very unstable results, since in some cases extremely homogeneous numbers of individuals live in an address class. Second, using indirect estimates, i.e. using the regression information on higher than SMP level did not show significant differences in the quality of the estimates. With regards to the importance of objective 1, the community separate regression estimator was preferred for SMP 0 and 1.
- **EBLUP** The classical Battese-Harter-Fuller unit-level estimator (Battese et al., 1988) was considered as the main small area estimator.
- **Weighted EBLUP (YOURAO)** An extension of the EBLUP using design weights was proposed by You and Rao (2002). This estimator also fulfils the necessary benchmarking conditions to aggregate the small area estimates to the design-based national estimate.

In all cases where auxiliary variables could be included, the necessary demographic variables from the population register were applied, i.e. number of persons, gender, and age classes. In the census test in 2001, the correlation between register counts and real counts was estimated to the level of 0.993 (cf. Münnich et al., 2012a, p. 70) and, thus, the register count is a very efficient auxiliary variable in terms of objective 1.

The Fay and Herriot (1979) basic area-level estimator was also considered. However, due to the very highly correlated population register information, this estimator was generally outperformed by the unit-level estimator and, hence, was omitted in this overview. As well as normal distribution-based models, estimators based on the binomial or Poisson distributions have been applied that account for the count structure. These estimators are built on the best prediction (BP) approach of Jiang and Lahiri (2001) with a setup similar to the one used in González-Manteiga et al. (2007) and Münnich et al. (2009). The estimation was done based on the *R*-package *lme4*. Details can be found in the given references.

For reasons of coherence, we focus in the next section on the main findings on the impact of sampling designs and some selected results in terms of objective 2. Some additional results of the project are as follows:

- For objective 1, the community-separate GREG estimator yielded convincing results which could not be outperformed by small area estimators. The main reason is that the SMPs of type 0 and 1 are not sufficiently small, so that model-based methods cannot show their advantage. Additionally, accuracy estimation is much easier when applying design-based methods.
- Objective 2 is much more complicated. Here, in many cases the YOURAO estimator was the best solution. However, it seems very important to think further about additional sources of auxiliary information in the future in order to further improve model-based estimates. The information in the population registers in many cases is not very efficient.
- Further research needs to be done when there is interest in deriving high-dimensional tables or the *one-number-census*. A generalized calibration routine is under development, which at least allows implementing hierarchical information on areas and domains with different penalties.

4. Estimation results

Within the census sampling and estimation research project, a large number of Monte Carlo simulations have been conducted using sampling from the register dataset. This dataset was synthetically enlarged by some objective 2 variables using other sources like the Microcensus so that the final dataset was close to reality. The procedure is described in Münnich et al. (2012a) and Kolb (2013).

4.1. Results for classic estimation

In an early stage of the project, different sampling approaches have been evaluated using a classical design-based simulation study. As main measure for the evaluation the RRMSE was applied using the known true value from the above mentioned universe. The results for three estimators under different sampling designs are shown in Figure 2, i.e. the HT, the GREG, the EBLUP.

Every grid in Figure 2 shows the results for one estimator for the federal state of Saarland. As auxiliary variable, the register counts were used. The sampling designs are described in Table 2. The abbreviation BV in Table 2 denotes balancing variables in the case of balanced sampling (cf. Tillé, 2011). These were the register variables address size class (ADC) or nationality (NAT).

The different sampling designs are presented in rows. For each sampling design three rows of ticks are presented, which denote the RRMSEs of 52 communities in Saarland. The upper tick covers one town, Saarbrücken. The middle ticks denote the RRMSEs of the large communities above 10,000 inhabitants. And finally, the bottom ticks present the RRMSEs of the smaller communities. The long line yields the kernel density estimates of the RRMSEs from all communities.

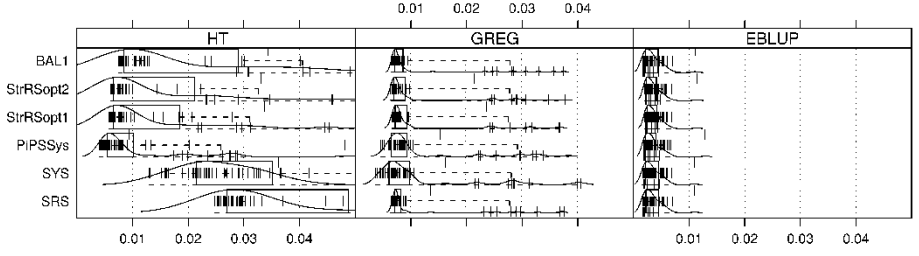


Figure 2. Comparisons of RRMSEs for various sampling designs for three total population estimates in Saarland

The names of the sampling designs in Table 2 refer to the classical designs within the SMPs. The allocation between the communities was drawn proportionally to the number of addresses.

Table 2. Different sampling designs - acronyms and their meanings

Acronym	description
BAL1	Balanced sampling; BV: ADC, NAT
StrRSopt2	Stratified random sample under optimal allocation (addresses)
StrRSopt1	Stratified random sample under optimal allocation (persons)
PiPSSys	π-PS systematic random sample
SYS	Systematic random sample
SRS	Simple random sample

As one can see in Figure 2, the EBLUP is fairly robust against the given sampling designs since the register counts allow for a strong model. However, one has to note that here no Neyman-based stratified design was applied, which would already negatively influence the results for the EBLUP. In terms of design-based methods, the design had considerable impact on the accuracy of the estimators. Simple random sampling, as expected, is very inefficient using the HT and much better for the GREG. Since the PiPSSys selects almost all large addresses, which seems hardly preferably in terms of a representative census sample, the accuracy of the results was considerably dependent on the address size structure of the communities. Balanced sampling did not yield very good results since the balancing variables seemed not to be so powerful and not many variables were available from the population register. The easier to implement stratified sampling designs seemed preferable with respect to accuracy, simplicity, and robustness against changes of settings as long as no specific allowance had to be made for smaller communities.

Finally, the results pointed to the use of stratified random sampling with some further optimization. The main gain in efficiency was stratification by address size. By the reasons given above, a box-constraint optimal allocation was introduced that guaranteed the necessary efficiency while still avoiding too much variation in the weights. Further, in any area, the census estimates were considerably more efficient than Microcensus estimates and no sub-population was drawn with a probability greater than 50%.

The results for the RRMSE under stratified sampling and box-constraint optimal allocation for different federal states are shown in Figure 3. The ordinate displays the different types of sampling points (from zero to three), while the abscissa indicates the RRMSE of the GREG estimator. Note that the results are theoretical results based on the address structure within the register and on the preassigned correlation of 0.993 between the register and true counts of people within addresses. Figure 3 shows that a-priori accuracy goals given in the census law were met in all SMPs of type 0 and 1, except for one community which failed slightly.

One has to note that even if the accuracy within SMPs 2 and 3 seems much lower, most SMPs of type 2 are still under 1% RRMSE which would have been the theoretical quality threshold using the objective 2 definition. Münnich et al. (2012a) showed that aggregating several estimates yields a RRMSE which is at least as good as the worst of the separate areas, which guarantees a hierarchical improvement by aggregation. Within the simulations, it turned out that this improvement, in general, is considerable.

Since the estimates of the total population (objective 1) were expected to be used for fiscal equalisation schemes between federal states and communities, special attention had to be paid to the estimator. Different possibilities were available for the estimation of the β -parameter of the GREG. Finally, the β -coefficient was estimated separately by SMP-type level, which in terms of the census law was separately by large community-level (above 10,000 inhabitants).

This was important to avoid a consideration of quality effects using indirect estimates employing information from other areas.

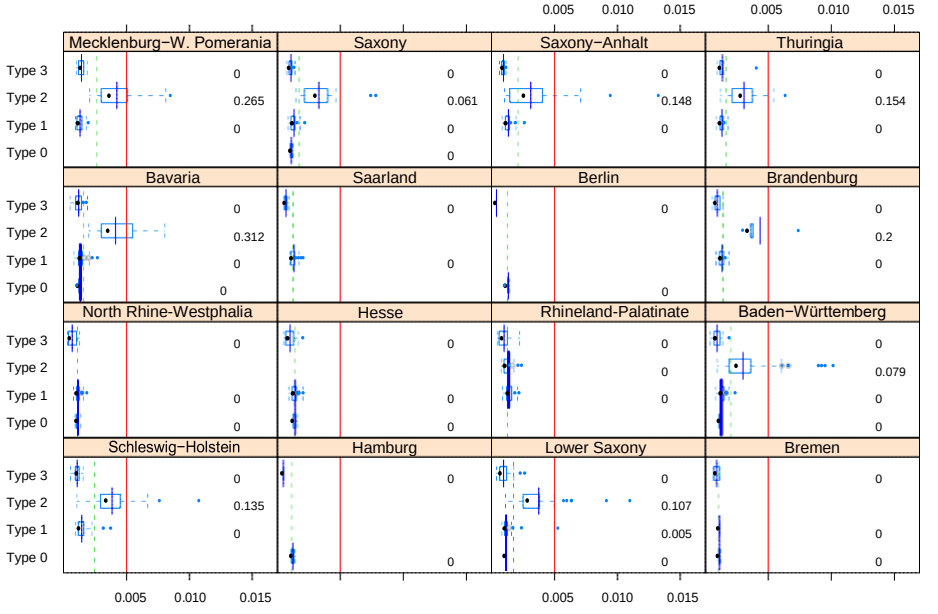


Figure 3. RRMSE of GREG estimates of population totals by federal state and sampling point type

4.2. Further ideas on small area applications

One of the objective 2 variables in the research project was the number of persons with foreign nationality (here the Turkish population) who had moved to Germany within a certain time span. Different estimators were applied to this estimation problem and the resulting RRMSEs are depicted in Figure 4 (Münnich et al. 2012a, pp. 104).

The descriptions of the headings in Figure 4 are provided in Table 3. In this Figure it is clear that YOURAO and EBLUP achieved the best results. However, for type 3 SMPs, the YOURAO estimator still yielded slightly better results, especially for earlier years of interest. Amazingly, the small area estimators also performed well in most cases of larger areas of type 0 and 1.

As a very important task in small area modeling, we have to consider vertical coherence, i.e. the aggregated small area estimates shall sum up to the national level estimates. It is well known that the GREG and YOURAO estimators fulfil this benchmarking condition. However, as a slightly more detailed assumption, we consider coherence to the next level, which in the German Census should also

include coherence at the district level. This is, therefore, measured as the difference between the sum of the lower level total estimates and the higher total estimate.

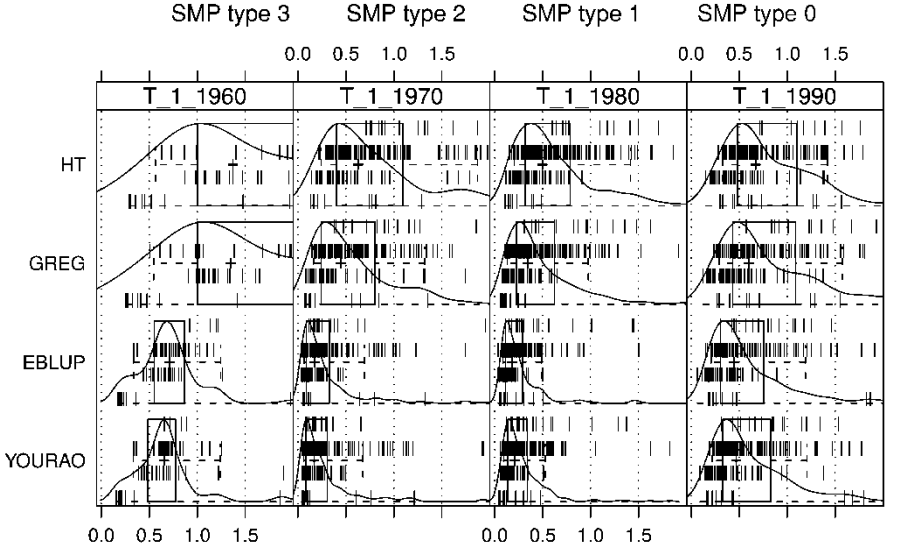


Figure 4. RRMSE for the estimation of selected years of moving in

Figure 5 indicates that the GREG is always coherent as long as the regression coefficient was estimated on the upper level which was the case in this example. The EBLUP suffers considerably from a lack in coherence. The YOURAO shows little deviations which may in fact be a result of a lack of the model in some districts. Further, the benchmarking condition holds only for the level on the β estimates, which here was the federal state level, which is higher than the district level. However, the deviation from perfect coherence is already small and much better than in the case of the EBLUP.

Table 3. Description of headings in Figure 4

Heading	description
	Number of persons, with Turkish nationality which
T_1960	moved to Germany between 1950 and 1960
T_1970	moved to Germany between 1960 and 1970
T_1980	moved to Germany between 1970 and 1980
T_1990	moved to Germany between 1980 and 1990

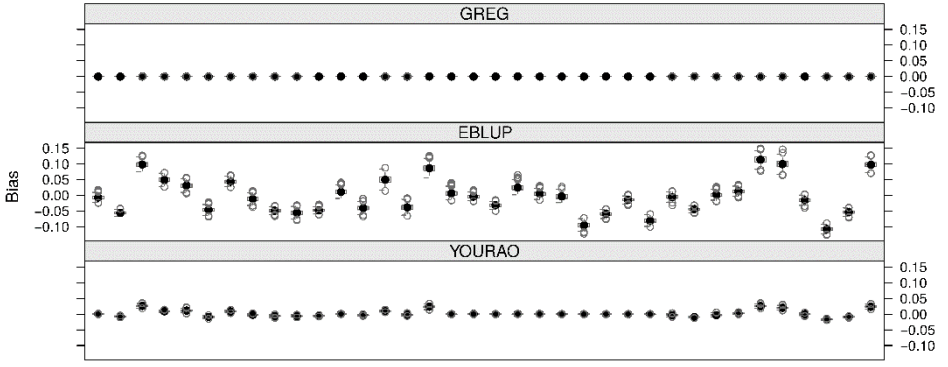


Figure 5. Coherence of aggregated SMPs estimates against district level estimates

To further force coherence of small area methods using different types of estimator the use of an extended calibration functional with penalties for further constraints based on regional small area estimates could be applied. By legal reasons, the first objective was exactly met on SMP level (0 and 1). Selected objective 2 estimates were met with high precision on the district level, whereas a lower preassigned precision was reached by other objective 2 estimates on SMP level. The variation of weights was constrained within the Gelman-bounds. It was possible to control the penalties separately on different levels and outcomes of covariates. Small area estimates for the totals can be used as benchmarks for objective 2 estimation. It is possible to apply different small area methods like, for example, the Battese-Harter-Fuller (Battese et al., 1988), the You-Rao (You and Rao, 2002), the Fay-Herriot and other estimators. The Lagrange multipliers provide a means to understand possible strains on area, domain or outcome of variables. This generalized calibration routine can be drawn from Münnich et al. (2012c) or Wagner (2013).

For a deeper overview of results from the entire study, we refer to Münnich et al. (2012a). The results suggested that the design recommendation still left enough space for applying small area methods. However, if a wider set of auxiliary variables was available from registers, e.g. by using matching methods, we would expect still a considerable improvement in the small area estimators.

5. Conclusions

As an outcome of the census sampling and estimation research project on the first German register-assisted census a recommendation was made for adopting a hierarchical SMP structure and a box-constraint optimal allocation for the sample sizes of addresses. For the first objective the use of a SMP-separate GREG was suggested. Either GREG or YOURAO estimators seemed adequate for the second objective depending on the target variable. An important consideration was that

objective 1 estimates were used to construct new population figures. Furthermore, the coherence of estimates was a very important target. In the case of a mix of methods on different hierarchies, an application of the generalized calibration method may be considered in the future.

It had to be ensured that the chosen methods were computationally tractable. Multinomial small area estimates may be promising to be applied in the future but currently suffer from the computational effort. To achieve further improvements of model-based estimators, the use of linking and matching of several registers should be further analysed, e.g. using specialised matching routines.

More information about the German Census 2011 can be found on the official website www.zensus2011.de or in Münnich et al. (2012a).

Acknowledgements

We would like to thank the German Federal Statistical Office and the Statistical Offices of the Federal States for the opportunity to carry out the underlying research within the Census sampling and estimation research project. Further, we thank Professor Li-Chun Zhang for providing the opportunity to present the work in the invited session on census small area applications during the SAE 2014 in Poznan. Finally, special thanks go to the editors and reviewers who helped to improve the readability of the text considerably.

REFERENCES

BATTESE, G. E., HARTER, R. M., 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), pp. 28–36.

BECHTOLD, S., (2013). The new register-based census of Germany − a multiple source mixed mode approach. In: Presentation on the 59th World Statistics Congress (WSC), Hong Kong, August 2013,
URL http://www.statistics.gov.hk/wsc/IPS027-P2-S.pdf.

FAY, R. E., 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, Vol. 74, No. 366, pp. 269–277.

FRIEDRICH, U., MÜNNICH, R., DE VRIES, S., WAGNER, M., (2015). Fast integer-valued algorithms for optimal allocations under constraints in stratified sampling. Resubmit- ted.

GABLER, S., GANNINGER, M., MÜNNICH, R., (2012). Optimal allocation of the sample size to strata under box constraints. Metrika 75(2), pp. 151–161.

GELMAN, A., (2007). Struggles with survey weighting and regression modeling. Statistical Science, pp. 153–164.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D., SANTAMARÍA, L., (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. Comput Stat Data Anal 51:2720–2733, DOI 10. 1016/j.csda.2006.01.012.

GUIBLIN, P., LONGFORD, N., HIGGINS, N., (2004). Standard estimators for small areas: Sas programs and documentation. Tech. rep., EURAREA − IST-2000-26290.

JIANG, J., LAHIRI, P., (2001). Empirical best prediction for small area inference with binary data. Annals of the Institute of Statistical Mathematics 53, pp. 217–243, dOI 10.1023/A:1012410420337.

KLEBER, B., MALDONADO, A., SCHEUREGGER, D., ZIPRIK, K., (2009). Aufbau des anschriften- und gebäuderegisters für den zensus 2011. Wirtschaft und Statistik 7, pp. 629–640.

KOLB, J. P., (2013). Methoden zur erzeugung synthetischer simulationsgesamtheiten. PhD thesis, Universität Trier.

MÜNNICH, R., MAGG, K., SOSTRA, K., SCHMIDT, K., WIEGERT, R., (2004). Workpackage 10: Variance estimation for small area estimates: Deliverables 10.1 and 10.2. URL http://www dacseis de-IST-2000-26057-DACSEIS Reports.

MÜNNICH, R., BURGARD, J. P., VOGT, M., (2009). Small area estimation for population counts in the German Census 2011. In: Proceedings of the Joint Statistical Meeting of the American Statistical Association. Washington.

MÜNNICH, R., BURGARD, P., VOGT, M., (2009). Small area estimation for population counts in the German Census 2011. In: Section on Survey Research Methods JSM 2009.

MÜNNICH, R., GABLER, S., GANNINGER, M., BURGARD, J. P., KOLB, J. P., (2012a). Stichprobenoptimierung und Schätzung im Zensus 2011. Statistisches Bundesamt.

MÜNNICH, R., SACHS, E. W., WAGNER, M., (2012b). Numerical solution of optimal allocation problems in stratified sampling under box constraints. AStA Advances in Statistical Analysis 96 (3), pp. 435–450.

MÜNNICH, R., WAGNER, M., SACHS, E. W., (2012c). Calibration benchmarking for small area estimates: An application to the German Census 2011. Symposium on the Anal- ysis of Survey Data and Small Area Estimation in Honour of the 75th Birthday of J. N. K Rao.

RAO, J. N. K., (2003). Small Area Estimation. Wiley series in survey methodology, John Wiley and Sons, New York.

THE EURAREA CONSORTIUM, (2004). Project reference volume vol. 2: Explanatory appendices. Tech. rep., EURAREA − IST-2000-26290.

TILLÉ, Y., (2011) Sampling algorithms. Springer.

WAGNER, M., (2013). Numerical optimization in survey statistics. PhD thesis, Univer- sität Trier, Universitätsring 15, 54296 Trier.

YOU, Y., RAO, J. N. K., (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. The Canadian Journal of Statistics / La Revue Canadienne de Statistique 30(3), pp. 431–439.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 41–66

A COMPARISON OF SMALL AREA ESTIMATION METHODS FOR POVERTY MAPPING

María Guadarrama[1], **Isabel Molina**[2], **J. N. K. Rao**[3]

ABSTRACT

We review main small area estimation methods for the estimation of general non-linear parameters focusing on FGT family of poverty indicators introduced by Foster, Greer and Thorbecke (1984). In particular, we consider direct estimation, the Fay-Herriot area level model (Fay and Herriot, 1979), the method of Elbers, Lanjouw and Lanjouw (2003) used by the World Bank, the empirical Best/Bayes (EB) method of Molina and Rao (2010) and its extension, the Census EB, and finally the hierarchical Bayes proposal of Molina, Nandram and Rao (2014). We put ourselves in the point of view of a practitioner and discuss, as objectively as possible, the benefits and drawbacks of each method, illustrating some of them through simulation studies.

Key words: area level model, non-linear parameters, empirical best estimator, hierarchical Bayes, poverty mapping, unit level models.

1. Introduction

Poverty maps are an important source of information on the regional distribution of poverty and are currently used to support regional policy making and to allocate funds to local jurisdictions. Good examples are the poverty and inequality maps produced by the World Bank for many countries all over the world. In the U.S., the Small Area Income and Poverty Estimates (SAIPE) program (http://www.census.gov/hhes/www/saipe) of the Census Bureau provides annual estimates of income and poverty statistics for all school districts, counties, and states, for the administration of federal, state and local programs and the allocation of federal funds to local jurisdictions. In Europe, the joint project "Poverty Mapping in the New Member States of the European Union" between the World Bank and the European Commission was aimed to construct poverty maps for the new members of the EU.

[1]Department of Statistics, Universidad Carlos III de Madrid. Address: C/Madrid 126, 28903 Getafe (Madrid), Spain, Tf: +34 916249859. E-mail: maria.guadarrama@uc3m.es

[2]Department of Statistics, Universidad Carlos III de Madrid. Address: C/Madrid 126, 28903 Getafe (Madrid), Spain, Tf: +34 916249887. E-mail: isabel.molina@uc3m.es

[3]School of Mathematics and Statistics, Carleton University. E-mail: jrao@math.carleton.ca

The TIPSE (The Territorial Dimension of Poverty and Social Exclusion in Europe) project, commissioned by the European Observation Network for Territorial Development and Cohesion (ESPON) program, aims to support policy by creating a regional database and associated maps of poverty and social exclusion indicators. In Mexico, the National Council for the Assessment of the Social Development Policy (CONEVAL) is committed by law to produce regular poverty and inequality estimates at the state level by population subgroups and at municipality level.

Obtaining accurate poverty maps at high levels of disaggregation is not straightforward because of insufficient sample size of official surveys in some of the target regions. Direct estimates, obtained with the region-specific sample data, are unstable in the sense of having very large sampling errors for regions with small sample size. Very unstable poverty estimates might make the seemingly poorer regions in one period appear as the richer in the next period, which can be contradictory. On the other hand, very stable but biased estimates (e.g., too homogeneous across regions) might make identification of the poorer regions difficult.

Here we review the main methods for the estimation of general non-linear small area parameters, focusing for illustrative purposes on a specific family of poverty indicators introduced in Section 2. Specifically, in Section 3 we describe direct estimation, the EBLUP based on the Fay-Herriot area level model (Fay and Herriot, 1979), the method of Elbers, Lanjouw and Lanjouw (2003), the empirical Best/Bayes (EB) method of Molina and Rao (2010) together with its variation called Census EB, and hierarchical Bayes (HB) method of Molina, Nandram and Rao (2014). We discuss advantages and disadvantages of each procedure from a practical point of view. In Section 4 we illustrate their performance in simulations under several scenarios, including the cases of informative sampling or the presence of outliers. Finally, in Section 5 we draw some conclusions.

2. Poverty indicators

In this paper, we will focus on the FGT family of poverty indicators introduced by Foster, Greer and Thorbecke (1984). Consider a population P of size N that is partitioned into D domains or areas $P_1,\ldots,P_D$, of sizes $N_1,\ldots,N_D$. Let E_{di} be a measure of welfare for individual i ($i=1,\ldots,N_d$) in area d ($d=1,\ldots,D$). Let z be the poverty line, that is, the value such that when $E_{di}<z$, individual i from area d is regarded as “at risk of poverty". Then, the FGT family of poverty indicators for area d is given by

$$F_{\alpha d}=\frac{1}{N_d}\sum_{i=1}^{N_d}\left(\frac{z-E_{di}}{z}\right)^{\alpha} I(E_{di}<z),\quad \alpha\geq 0, d=1,\ldots,D, \tag{1}$$

where $I(E_{di} < z) = 1$ if $E_{di} < z$, and $I(E_{di} < z) = 0$ otherwise. For $\alpha = 0$ we obtain the proportion of individuals “at risk of poverty", that is, the poverty incidence or at-risk-of-poverty rate. For $\alpha = 1$, we get the average of the relative distances to not being “at risk of poverty", called the poverty gap. The poverty incidence measures the frequency of poverty, whereas the poverty gap measures the intensity of poverty. We remark that the unit level methods introduced in this paper can be applied to estimate any desired population characteristic that is obtained as a real measurable function of a continuous variable, as long as this variable follows the considered model in each method.

3. Estimators

Estimation of population characteristics is typically based on a sample s drawn from the population P. We denote by $s_d = s \cap P_d$ the subsample from area d of size $n_d < N_d$ and by $r_d = P_d - s_d$ the complement of s_d, of size $N_d - n_d$. The overall sample size is $n = n_1 + \cdots + n_D$. The following subsections describe common estimators of poverty indicators obtained from the sample data.

3.1. Direct estimators

Turning now to estimation in a given domain or area d, a direct estimator is an estimator obtained using only the n_d observations from that area, provided that this area has been sampled (i.e., $n_d > 0$). The FGT poverty indicator (1) of order α for area d can be expressed as a linear parameter as follows

$$F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha di}, \quad F_{\alpha di} = \left(\frac{z - E_{di}}{z} \right)^{\alpha} I(E_{di} < z), \quad i = 1, \ldots, N_d.$$

Then, the basic direct estimator of $F_{\alpha d}$ is simply given by

$$\hat{F}_{\alpha d}^{\mathrm{DIR}} = N_d^{-1} \sum_{i \in s_d} w_{d,i} F_{\alpha di}, \tag{2}$$

where $w_{d,i} = \pi_{d,i}^{-1}$ is the sampling weight of unit i from area d and $\pi_{d,i}$ is the inclusion probability of unit i in the subsample s_d.

Below we list the advantages and disadvantages of direct estimators, such as (2), for small area estimation.

Advantages:

- They are (at least approximately) design-unbiased and design-consistent (as

$n_d \to \infty$). Thus, they perform well under complex sampling designs, including informative sampling, as long as they are calculated using the correct inclusion probabilities.

- They do not require model assumptions; that is, they are completely nonparametric.

Disadvantages:

- They are very inefficient for areas with very small n_d.
- They cannot be calculated for nonsampled areas (i.e., with $n_d = 0$).

3.2. Fay-Herriot model

Fay-Herriot (FH) area level model links the parameters of interest for all the areas, $F_{\alpha d}$, $d = 1,\dots,D$, through a linear model as

$$F_{\alpha d} = \mathbf{x}_d'\beta + u_d, \quad d = 1,\dots,D, \tag{3}$$

where $\mathbf{x}_d$ is a p-vector of area level covariates, β is the regression parameter common for all areas, and u_d is the area-specific regression error, also called random effect for area d. We assume that area random effects u_d are independent and identically distributed (iid), with unknown variance σ_u^2, that is, $u_d \overset{iid}{\sim} (0, \sigma_u^2)$. Note that true values $F_{\alpha d}$ are not observable and therefore model (3) cannot be directly fitted. However, we can make use of a direct estimator $\hat{F}_{\alpha d}^{\mathrm{DIR}}$ of $F_{\alpha d}$. FH model assumes that $\hat{F}_{\alpha d}^{\mathrm{DIR}}$ is design-unbiased, with

$$\hat{F}_{\alpha d}^{\mathrm{DIR}} = F_{\alpha d} + e_d, \quad d = 1,\dots,D, \tag{4}$$

where e_d is the sampling error for domain d. We assume that sampling errors e_d are independent of random effects u_d and satisfy $e_d \overset{ind}{\sim} (0, \psi_d)$, where the sampling variances ψ_d, $d = 1,\dots,D$, are assumed to be known. Combining (3) and (4), we obtain a linear mixed model

$$\hat{F}_{\alpha d}^{\mathrm{DIR}} = \mathbf{x}_d'\beta + u_d + e_d, \quad d = 1,\dots,D. \tag{5}$$

The best linear unbiased predictor (BLUP) of $F_{\alpha d} = \mathbf{x}_d'\beta + u_d$ under model (5) is given by

$$\tilde{F}_{\alpha d}^{\mathrm{FH}} = \mathbf{x}_d'\tilde{\beta} + \tilde{u}_d, \tag{6}$$

where $\tilde{u}_d = \gamma_d(\hat{F}_{\alpha d}^{\mathrm{DIR}} - \mathbf{x}_d'\tilde{\beta})$ is the BLUP of u_d, with $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$ and where $\tilde{\beta}$ is the weighted least squares estimator of β, given by

$$\tilde{\beta} = \left(\sum_{d=1}^{D} \gamma_d \mathbf{x}_d \mathbf{x}_d'\right)^{-1} \sum_{d=1}^{D} \gamma_d \mathbf{x}_d \hat{F}_{\alpha d}^{\mathrm{DIR}}.$$

In practice, the variance σ_u^2 of the area effects u_d is unknown and needs to be estimated. Common estimation methods are maximum likelihood (ML) and restricted maximum likelihood (REML). REML corrects for the degrees of freedom due to estimating β and leads to a less biased estimator of σ_u^2 for finite sample size n. Let $\hat{\sigma}_u^2$ be the resulting estimator. Replacing $\hat{\sigma}_u^2$ for σ_u^2 in (6), we obtain the empirical BLUP (EBLUP) of $F_{\alpha d}$, denoted here as $\hat{F}_{\alpha d}^{\mathrm{FH}}$ and called hereafter FH estimator.

A second-order correct estimator of MSE ($\hat{F}_{\alpha d}^{\mathrm{FH}}$) is given in Rao (2003, Chapter 7), assuming normality of u_d and e_d. Good and bad properties of FH estimator (6) are listed below, including particular properties for poverty mapping.

Advantages:

- The BLUP under FH model can be expressed as a weighted combination of the direct and the regression-synthetic estimators, that is,

$$\tilde{F}_{\alpha d}^{\mathrm{FH}} = \gamma_d \hat{F}_{\alpha d}^{\mathrm{DIR}} + (1-\gamma_d)\mathbf{x}_d'\tilde{\beta}, \quad d = 1, \ldots, D. \tag{7}$$

  with weight $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$. Then, for an area d in which the direct estimator $\hat{F}_{\alpha d}^{\mathrm{DIR}}$ is inefficient, that is, with a large sampling variance ψ_d compared to the unexplained between-area variability σ_u^2, γ_d becomes small and $\tilde{F}_{\alpha d}^{\mathrm{FH}}$ borrows more strength from the other areas through the regression-synthetic estimator $\mathbf{x}_d'\tilde{\beta}$. On the other hand, for an area d in which the direct estimator $\hat{F}_{\alpha d}^{\mathrm{DIR}}$ is efficient, that is, with small sampling variance ψ_d compared to the unexplained between-area variability σ_u^2, γ_d is large and $\tilde{F}_{\alpha d}^{\mathrm{FH}}$ attaches more weight to the direct estimator. Thus, FH estimator automatically borrows strength for the areas where it is needed.
- If $\gamma_d > 0$ for area d, it makes use of the sampling weights $w_{d,i}$ through the direct estimator $\hat{F}_{\alpha d}^{\mathrm{DIR}}$. Thus, it is design-consistent (as $n_d \to \infty$). As a consequence, it is less affected by informative sampling provided that the direct estimator is calculated using the correct inclusion probabilities.
- Due to the aggregation of data, it is not very much affected by isolated unit level outliers.
- It requires only area level auxiliary information and therefore avoids the confidentiality issues associated with micro-data.

Disadvantages:

- The sampling variances ψ_d are assumed to be known, but in practice they are estimated. It is not easy to incorporate the uncertainty due to estimation of the sampling variances in the MSE.
- The number of observations used to fit the FH model is the number of areas D, which is typically much smaller than the number of observations used to fit unit level models, n. Thus, model parameters are estimated with less efficiency and therefore the efficiency gains with respect to direct estimators are expected to be smaller than under unit level models.
- It requires normality of u_d and e_d for MSE estimation. This might not hold for very complex poverty indicators.
- If we want to estimate several indicators depending on a common continuous variable, it requires separate modeling and searching for good covariates for each indicator.
- Once the model is fitted at the area level, small area estimates $\hat{F}_{\alpha d}^{\mathrm{FH}}$ cannot be further disaggregated for subdomains or subareas within the areas unless a new good model is found at that subarea level.

3.3. ELL method

The method of Elbers, Lanjouw and Lanjouw (2003), called hereafter ELL method, assumes a unit level linear mixed model for a log-transformation of the variable measuring welfare of individuals, with random effects for the sampling clusters or primary sampling units. For comparability with the rest of the methods presented here, in the following we assume that the sampling clusters are the areas. In this case, the model becomes the nested error model of Battese, Harter and Fuller (1988) for the log-transformation of the welfare variables, that is, $Y_{di} = \log(E_{di})$ is assumed to be linearly related with a p-vector of auxiliary variables $\mathbf{x}_{di}$, which may include unit-specific and area-specific covariates, and includes random area effects u_d as follows

$$Y_{di} = \mathbf{x}_{di}'\beta + u_d + e_{di}, \quad i = 1,\ldots,N_d, d = 1,\ldots,D. \tag{8}$$

Here, β is a p-vector of regression coefficients, $u_d \overset{iid}{\sim} (0,\sigma_u^2)$, $e_{di} \overset{ind}{\sim} (0,\sigma_e^2 k_{di}^2)$, where u_d and e_{di} are independent and k_{di} are known constants.

ELL estimator of $F_{\alpha d}$ is given by the marginal expectation $\hat{F}_{\alpha d}^{\mathrm{ELL}} = E[F_{\alpha d}]$ under model (8). This estimator and its MSE are approximated by a bootstrap method. In this bootstrap procedure, random effects u_d^* and model errors e_{di}^* are generated from

residuals obtained by fitting model (8) to survey data. Then, a bootstrap census of Y-values is generated as

$$Y_{di}^{*} = \mathbf{x}_{di}'\hat{\boldsymbol{\beta}} + u_d^{*} + e_{di}^{*}, \quad i = 1,\ldots,N_d,\, d = 1,\ldots,D,$$

where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$. The generation is repeated for $a = 1,\ldots,A$, obtaining A censuses. Then, for each bootstrap census a, the FGT poverty indicator for area d is calculated as

$$F_{\alpha d}^{*(a)} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left(\frac{z-\exp(Y_{di}^{*(a)})}{z}\right)^{\alpha} I(\exp(Y_{di}^{*(a)}) < z).$$

The ELL estimator of $F_{\alpha d}$ is then approximated by averaging over the A generated censuses, that is,

$$\hat{F}_{\alpha d}^{\mathrm{ELL}} = \frac{1}{A}\sum_{a=1}^{A} F_{\alpha d}^{*(a)}.$$

The MSE of $\hat{F}_{\alpha d}^{\mathrm{ELL}}$ is then estimated as follows

$$\mathrm{mse}(\hat{F}_{\alpha d}^{\mathrm{ELL}}) = \frac{1}{n_d}\sum_{a=1}^{A}(F_{\alpha d}^{*(a)} - \hat{F}_{\alpha d}^{\mathrm{ELL}})^2.$$

Advantages and disadvantages of ELL method are listed below.

Advantages:

- It is based on unit level data, which are richer than area level data and sample size is much larger (n compared to D).
- ELL method can be applied to estimate general indicators defined as a function of the model response variables Y_{di}.
- They are model-unbiased if the model parameters are known.
- Once the model is fitted, estimates can be obtained at whatever subarea level.

Disadvantages:

- In terms of model MSE, ELL estimates perform poorly and can even perform worse than direct estimators when unexplained between-area variation is significant, see Molina and Rao (2010). In fact, for the estimation of domain means, ELL estimates are basically equal to regression-synthetic estimators, which assume the regression model without further between-area variation.
- They are based on a model assumption. Hence, model checking is crucial.

- They are not design-unbiased and can be seriously biased under informative sampling.
- They can be seriously affected by unit level outliers.
- If cluster effects are included in the model instead of area effects, but area effects are significant, ELL estimates of the model MSE can seriously underestimate the true MSE. Even if area effects are included in the model, ELL estimates of MSE do not track correctly the true MSE for each area.

3.4. Empirical Best/Bayes EB method

The empirical Best/Bayes (EB) method of Molina and Rao (2010) assumes that the population variables Y_{di} follow the nested error model (8) with normality of random effects u_d and errors e_{di}. Under that model, the area vectors $\mathbf{Y}_d = (Y_{d1},\ldots,Y_{dN_d})'$ are independent for $d = 1,\ldots,D$ and satisfy $\mathbf{Y}_d \overset{ind}{\sim} N(\mu_d, \mathbf{V}_d)$, where $\mu_d = \mathbf{X}_d\beta$ and $\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}'_{N_d} + \sigma_e^2 \mathbf{A}_d$, for $\mathbf{A}_d = \mathrm{diag}(k_{di}^2; i = 1,\ldots,N_d)$. For an area parameter $\delta_d = h(\mathbf{Y}_d)$, the estimator that minimizes the MSE, called best estimator, is given by

$$\hat{\delta}_d^{\mathrm{B}} = E_{\mathbf{Y}_{dr}}[h(\mathbf{Y}_d)|\mathbf{Y}_{ds};\theta] = \int h(\mathbf{Y}_d) f(\mathbf{Y}_{dr}|\mathbf{Y}_{ds};\theta) d\mathbf{Y}_{dr}, \tag{9}$$

where $f(\mathbf{Y}_{dr}|\mathbf{Y}_{ds};\theta)$ is the conditional distribution of the vector of out-of-sample values $\mathbf{Y}_{dr}$ in domain d given the sampled values $\mathbf{Y}_{ds}$ in that domain and θ is the vector of model parameters. Now replacing θ in (9) by an estimator $\hat{\theta}$, we get the empirical best (EB) estimator, $\hat{\delta}_d^{\mathrm{EB}}$.

Under the nested error model (8), the distribution of $\mathbf{Y}_{dr}|\mathbf{Y}_{ds}$ is easy to derive. First, we decompose $\mathbf{X}_d$ and $\mathbf{V}_d$ into sample and out-of-sample elements similarly as we do with $\mathbf{Y}_d$, that is,

$$\mathbf{Y}_d = \begin{pmatrix} \mathbf{Y}_{ds} \\ \mathbf{Y}_{dr} \end{pmatrix}, \quad \mathbf{X}_d = \begin{pmatrix} \mathbf{X}_{ds} \\ \mathbf{X}_{dr} \end{pmatrix}, \quad \mathbf{V}_d = \begin{pmatrix} \mathbf{V}_{ds} & \mathbf{V}_{dsr} \\ \mathbf{V}_{drs} & \mathbf{V}_{dr} \end{pmatrix}.$$

By the normality assumption, we have that $\mathbf{Y}_{dr}|\mathbf{Y}_{ds} \overset{ind}{\sim} N(\mu_{dr|s}, \mathbf{V}_{dr|s})$, where the conditional mean vector and covariance matrix are given by

$$\mu_{dr|s} = \mathbf{X}_{dr}\beta + \gamma_{dc}(\bar{y}_{dc} - \bar{\mathbf{x}}_{dc}^T\beta)\mathbf{1}_{N_d-n_d}, \tag{10}$$

$$\mathbf{V}_{dr|s} = \sigma_u^2(1-\gamma_d)\mathbf{1}_{N_d-n_d}\mathbf{1}_{N_d-n_d}^T + \sigma_e^2 \mathrm{diag}_{i\in r_d}(k_{di}^2). \tag{11}$$

Here, $\gamma_{dc} = \sigma_u^2(\sigma_u^2 + \sigma_e^2/c_{d\cdot})^{-1}$, for $c_{d\cdot} = \sum_{i\in s_d} c_{di}$ with $c_{di} = k_{di}^{-2}$, and $\bar{y}_{dc}$ and $\bar{\mathbf{x}}_{dc}$

are weighted sample means obtained as

$$\bar{y}_{dc} = \frac{1}{c_{d\cdot}} \sum_{i \in s_d} c_{di} Y_{di}, \quad \bar{\mathbf{x}}_{dc} = \frac{1}{c_{d\cdot}} \sum_{i \in s_d} c_{di} \mathbf{x}_{di}. \tag{12}$$

For complex non-linear parameters $\delta_d = h(\mathbf{Y}_d)$, the expectation given in (9) cannot be calculated analytically. In those cases, the EB estimator $\hat{\delta}_d^{\mathrm{EB}}$ is approximated by Monte Carlo. This requires to simulation of multivariate Normal vectors $\mathbf{Y}_{dr}^{(a)}$ of sizes $N_d - n_d$, $d = 1, \ldots, D$, from the (estimated) conditional distribution of $\mathbf{Y}_{dr}|\mathbf{Y}_{ds}$ and then to replication for $a = 1, \ldots, A$, which may be computationally unfeasible. Simulation of very large multivariate Normal vectors $\mathbf{Y}_{dr}^{(a)}$ can be avoided by noting that the conditional covariance matrix $\mathbf{V}_{dr|s}$, given by (11), corresponds to the covariance matrix of a random vector $\mathbf{Y}_{dr}^{(a)}$ generated from the model

$$\mathbf{Y}_{dr}^{(a)} = \mu_{dr|s} + v_d^{(a)} \mathbf{1}_{N_d - n_d} + \boldsymbol{\varepsilon}_{dr}^{(a)}, \tag{13}$$

where $v_d^{(a)}$ and $\boldsymbol{\varepsilon}_{dr}^{(a)}$ are independent and satisfy

$$v_d^{(a)} \sim N(0, \sigma_u^2(1-\gamma_d)) \quad \text{and} \quad \boldsymbol{\varepsilon}_{dr}^{(a)} \sim N(\mathbf{0}_{N_d - n_d}, \sigma_e^2 \mathrm{diag}_{i \in r_d}(k_{di}^2));$$

see Molina and Rao (2010). Using model (13), instead of generating a multivariate normal vector $\mathbf{Y}_{dr}^{(a)}$ of size $N_d - n_d$, we just need to generate $1 + N_d - n_d$ independent univariate normal variables $v_d^{(a)} \overset{ind}{\sim} N(0, \sigma_u^2(1-\gamma_d))$ and $\varepsilon_{di}^{(a)} \overset{ind}{\sim} N(0, \sigma_e^2 k_{di}^2)$, for $i \in r_d$. Then, we obtain the corresponding out-of-sample values $Y_{di}^{(a)}$, $i \in r_d$, from (13) using as means the corresponding elements of $\mu_{dr|s}$ given by (10). Using the vector $\mathbf{Y}_{dr}^{(a)}$ generated from (13), we construct the census vector $\mathbf{Y}_d^{(a)} = (\mathbf{Y}_{ds}', (\mathbf{Y}_{dr}^{(a)})')'$ and calculate the parameter of interest $\delta_d^{(a)} = h(\mathbf{Y}_d^{(a)})$. For a non-sampled area d (i.e., with $n_d = 0$), we generate $\mathbf{Y}_{dr}^{(a)}$ from (13) with $\gamma_{dc} = 0$ and in this case $\mathbf{Y}_d^{(a)} = \mathbf{Y}_{dr}^{(a)}$. The Monte Carlo approximation to the EB estimator (9) of $\delta_d = h(\mathbf{Y}_d)$ is then given by

$$\hat{\delta}_d^{\mathrm{EB}} \approx \frac{1}{A} \sum_{a=1}^{A} h(\mathbf{Y}_d^{(a)}). \tag{14}$$

In particular, to estimate the FGT poverty indicator given in (1), Molina and Rao (2010) assumed that $Y_{di} = T(E_{di})$ follow the nested error model (8), where E_{di} are variables measuring welfare and $T(\cdot)$ is a one-to-one transformation. In terms of the vector of transformed variables $\mathbf{Y}_d = (Y_{d1}, \ldots, Y_{dN_d})'$, the FGT poverty indicator

can be expressed as

$$F_{\alpha d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{z - T^{-1}(Y_{di})}{z} \right)^{\alpha} I(T^{-1}(Y_{di}) < z) = h_{\alpha}(\mathbf{Y}_d), \tag{15}$$

and the above EB method can be applied to the area parameter $\delta_d = h_{\alpha}(\mathbf{Y}_d)$.

In the case of complex parameters such as the FGT poverty indicators, analytic approximations for the MSE are hard to derive. Molina and Rao (2010) obtained a parametric boostrap MSE estimator following the bootstrap method for finite populations of González-Manteiga et al. (2008), see Molina and Rao (2010) for further details.

Note that both ELL and EB methods require a survey data file containing the observations from the target variable and the auxiliary variables, that is, $\{(Y_{di}, \mathbf{x}_{di}); i \in s_d,\ d = 1, \ldots, D\}$, and a census containing the values of the same auxiliary variables for all the units in the population, that is, $\{\mathbf{x}_{di}; i = 1, \ldots, N_d,\ d = 1, \ldots, D\}$. The EB method requires additionally the identification of the set of out-of-sample units r (or equivalently the sample units s) in the census P. Linking the survey and the census files is not always possible in practice. However, typically the area sample size n_d is really small compared to the population size N_d. Then, we can use the Census-EB estimator proposed by Correa, Molina and Rao (2012), and obtained by generating in each Monte Carlo replicate the full census vector $\mathbf{Y}_d$ rather than only the vector of out-of-sample observations $\mathbf{Y}_{dr}$. For this, we apply the Monte Carlo approximation (9) by generating $\mathbf{Y}_d^{(a)} = \mu_{d|s} + v_d^{(a)} \mathbf{1}_{N_d - n_d} + \varepsilon_d^{(a)}$, where $\mu_{d|s} = \mathbf{X}_d \beta + \gamma_{dc}(\bar{y}_{dc} - \bar{\mathbf{x}}_{dc}^T \beta) \mathbf{1}_{N_d}$ and $\varepsilon_d^{(a)} \sim N(\mathbf{0}_{N_d}, \sigma_e^2 \mathrm{diag}_{i=1,\ldots,N_d}(k_{di}^2))$. If the sampling fraction n_d/N_d is negligible, the Census-EB estimator of $\delta_d = F_{\alpha d}$ is practically the same as the original EB estimator.

Good properties and drawbacks of the EB method are listed below.

Advantages:

- It is based on unit level data, which are richer than the area level data and uses much larger sample size to fit the model.
- The EB method can be applied to estimate general indicators defined as functions of the response variables Y_{di}.
- Best estimators are model-unbiased.
- They are optimal in terms of minimizing the model MSE for known values of model parameters.
- EB estimates perform significantly better than ELL estimates when unexplained between-area variation is significant. For out-of-sample areas (with

$n_d = 0$), EB and ELL small area estimates are nearly the same. They are nearly the same for all areas if there is no unexplained between-area variation ($\sigma_u^2 = 0$).

- Once the model is fitted, estimates can be obtained at whatever subarea level.

Disadvantages:

- They are based on a model assumption. Hence, model checking is crucial.
- They are not approximately design-unbiased and can be seriously biased under informative sampling.
- They can be severely affected by unit level outliers.
- Parametric bootstrap estimates of the MSE of EB estimators are computationally intensive.

3.5. Hierarchical Bayes (HB) method

Computation of EB (and Census-EB) estimates supplemented with their MSE estimates is very intensive and might be unfeasible for very large populations or for very complex indicators. Note that to approximate the EB estimate by Monte Carlo, we need to construct a large number A of censuses $\mathbf{Y}^{(a)}$, where each one might be of huge size. Moreover, to obtain the parametric bootstrap MSE estimator, the Monte Carlo approximation needs to be repeated for each bootstrap replicate. Seeking for a computationally more efficient approach, Molina, Nandram and Rao (2014) developed the alternative hierarchical Bayes (HB) method for estimation of complex non-linear parameters. This approach does not require the use of bootstrap for MSE estimation because it provides samples from the posterior distribution, from which posterior variances play the role of MSEs, and any other useful posterior summary can be easily obtained.

The HB method is based on reparameterizing the nested error model (8) in terms of the intraclass correlation coefficient $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$ and considering priors for the model parameters $(\beta, \rho, \sigma_e^2)$ that reflect lack of knowledge. Concretely, the HB model is defined as

$$\text{(i)} \quad Y_{di}|u_d, \beta, \sigma_e^2 \overset{ind}{\sim} N(\mathbf{x}'_{di}\beta + u_d, \sigma_e^2 k_{di}^2), \quad i = 1, \ldots, N_d,$$

$$\text{(ii)} \quad u_d|\rho, \sigma_e^2 \overset{iid}{\sim} N\left(0, \frac{\rho}{1-\rho}\sigma_e^2\right), \quad d = 1, \ldots, D,$$

$$\text{(iii)} \quad \pi(\beta, \rho, \sigma_e^2) \propto \frac{1}{\sigma_e^2}, \quad \varepsilon \leq \rho \leq 1 - \varepsilon, \, \sigma_e^2 > 0, \beta \in \mathscr{R}^p,$$

where $\varepsilon > 0$ is chosen very small to reflect lack of knowledge. See the application carried out by Molina, Nandram and Rao (2014), where inference was not sensitive to a small change of ε.

The posterior distribution can be obtained in terms of posterior conditionals using the chain rule of probability as follows. First, note that under the HB approach, the random effects $\mathbf{u} = (u_1, \ldots, u_D)'$ are regarded as additional parameters. Then, the joint posterior pdf of the vector of parameters $\theta = (\mathbf{u}', \beta', \sigma_e^2, \rho)'$ given the sample values $\mathbf{Y}_s$ is given by

$$\pi(\mathbf{u}, \beta, \sigma_e^2, \rho | \mathbf{Y}_s) = \pi_1(\mathbf{u}|\beta, \sigma_e^2, \rho, \mathbf{Y}_s)\pi_2(\beta|\sigma_e^2, \rho, \mathbf{Y}_s)\pi_3(\sigma_e^2|\rho, \mathbf{Y}_s)\pi_4(\rho|\mathbf{Y}_s), \quad (16)$$

where the conditional pdfs $\pi_1, \ldots, \pi_3$ have known forms, but not π_4. However, since ρ is in a closed interval from $(0,1)$, we can generate values from π_4 using a grid method, for more details see Molina, Nandram and Rao (2014). Samples from $\theta = (\mathbf{u}', \beta', \sigma_e^2, \rho)'$ can then be generated directly from the posterior distribution in (16), avoiding the use of Markov Chain Monte Carlo (MCMC) methods. Under general conditions, a proper posterior distribution is guaranteed.

Given θ, population variables Y_{di} are all independent, satisfying

$$Y_{di}|\theta \overset{ind}{\sim} N(\mathbf{x}'_{di}\beta + u_d, \sigma_e^2 k_{di}^2), \quad i = 1, \ldots, N_d, \, d = 1, \ldots, D. \quad (17)$$

The posterior predictive density of $\mathbf{Y}_{dr}$ is then given by

$$f(\mathbf{Y}_{dr}|\mathbf{Y}_s) = \int \prod_{i \in r_d} f(Y_{di}|\theta)\pi(\theta|\mathbf{Y}_s)d\theta.$$

Finally, the HB estimator of a domain parameter $\delta_d = h(\mathbf{Y}_d)$ is given by

$$\hat{\delta}_d^{\mathrm{HB}} = E_{\mathbf{Y}_{dr}}(\delta_d|\mathbf{Y}_s) = \int h(\mathbf{Y}_d) f(\mathbf{Y}_{dr}|\mathbf{Y}_s)d\mathbf{Y}_{dr}. \quad (18)$$

The HB estimator can be approximated by Monte Carlo. For this, we first generate samples from the posterior $\pi(\theta|\mathbf{Y}_s)$. We generate a value $\rho^{(a)}$ from $\pi_4(\rho|\mathbf{Y}_s)$ using a grid method; then, a value $\sigma_e^{2(a)}$ is generated from $\pi_3(\sigma_e^2|\rho^{(a)}, \mathbf{Y}_s)$; next $\beta^{(a)}$ is generated from $\pi_2(\beta|\sigma_e^{2(a)}, \rho^{(a)}, \mathbf{Y}_s)$ and, finally, $\mathbf{u}^{(a)}$ is generated from $\pi_1(\mathbf{u}|\beta^{(a)}, \sigma_e^{2(a)}, \rho^{(a)}, \mathbf{Y}_s)$. This process is repeated a large number A of times to get a random sample $\theta^{(a)}$, $a = 1, \ldots, A$ from $\pi(\theta|\mathbf{Y}_s)$. Now for each generated value $\theta^{(a)}$ from $\pi(\theta|\mathbf{Y}_s)$, we generate the out-of-sample values $\{Y_{di}^{(a)}, i \in r_d\}$ from the distribution defined in (17). Thus, for each area d, we have generated an out-of-sample vector $\mathbf{Y}_{dr}^{(a)} = \{Y_{di}^{(a)}, i \in r_d\}$, and we have also the available sample data $\mathbf{Y}_{ds}$. Putting them together, we construct the full population vector $\mathbf{Y}_d^{(a)} = (\mathbf{Y}'_{ds}, (\mathbf{Y}_{dr}^{(a)})')'$.

Now using $\mathbf{Y}_d^{(a)}$, we compute the area parameter $\delta_d^{(a)} = h(\mathbf{Y}_d^{(a)})$. In the particular case of estimating an FGT poverty indicator, we have $\delta_d = F_{\alpha d} = h_\alpha(\mathbf{Y}_d)$ given in (15). Then, in Monte Carlo replicate a, we calculate $F_{\alpha d}^{(a)} = h_\alpha(\mathbf{Y}_d^{(a)})$. Finally, the HB estimator is approximated as

$$\hat{F}_{\alpha d}^{\mathrm{HB}} \approx \frac{1}{A}\sum_{a=1}^{A} F_{\alpha d}^{(a)}. \tag{19}$$

Benefits and deficiencies of HB method are listed below.

Advantages:

- It is based on unit level data, which are richer than area level data and uses much larger sample size to fit the model.
- HB method can be applied to estimate general indicators defined as function of the model response variables Y_{di}.
- HB estimators are model-unbiased.
- HB estimators are optimal in terms of minimizing the posterior variance.
- EB and HB methods are expected to give practically the same point estimates, see Molina, Nandram and Rao (2014). Thus, the proposed HB method has good frequentist properties.
- Once the model is fitted, estimates can be obtained at any subarea level.
- The proposed HB approach does not require the use of MCMC methods and therefore avoids the need of monitoring the convergence of Monte Carlo chains.
- Bootstrap methods for MSE estimation are not needed. Therefore, total computational time is considerably lower than in the EB method.
- Calculation of credible intervals or other posterior summaries is straightforward.

Disadvantages:

- It is based on model assumptions. Hence, model checking is crucial.
- HB estimators are not design-unbiased and can be seriously biased under informative sampling.
- HB estimators can be severely affected by unit level outliers.
- HB method is not directly extendable to more complex models without losing some of the mentioned advantages like avoiding MCMC.

4. Simulation studies

This section illustrates some of the mentioned advantages and drawbacks of the considered poverty mapping methods through simulation studies. Concretely, we will report results of simulations under three different scenarios: (i) Nested error model with simple random sampling. (ii) Nested error model with informative sampling. (iii) Nested error model with outliers.

Simulations were implemented in the statistical software environment R (R development core team 2013) using the package `lme4` (Bates et al. 2014), which fits Gaussian linear and nonlinear mixed-effects models, and the package `sae` (Molina and Marhuenda 2015), which contains functions for small area estimation, including calculation of direct, FH and EB estimates along with their MSE estimates.

4.1. Nested error model with simple random sampling

We consider the same model-based simulation setup as in Molina, Nandram and Rao (2014), where data are generated at the unit level following the nested error model (8). However, here we also include FH estimators derived from the FH area level model with the area means of the auxiliary variables as covariates. In addition, we include ELL and Census-EB estimators. The population is composed of $N = 20,000$ units, distributed in $D = 80$ areas with $N_d = 250$ units in each area. We consider two auxiliary variables X_1 and X_2 with known values for all the population units. Their values are generated as $x_{k,di} \sim \text{Bern}(p_{kd})$, $k = 1,2$, with success probabilities $p_{1d} = 0.3 + 0.5d/D$ and $p_{2d} = 0.2$, $d = 1,\ldots,D$. Response variables Y_{di} are generated from the nested error model (8) and the target variables are $E_{di} = \exp(Y_{di})$. The true values of the regression coefficients are $\beta = (3, 0.03, -0.04)'$. Variances of area effects and errors are taken as $\sigma_u^2 = 0.15^2$ and $\sigma_e^2 = 0.5^2$ respectively. The poverty line is set to $z = 12$, which is approximately 0.6 times the median of $\{E_{di}; i = 1,\ldots,N_d,\ d = 1,\ldots,D\}$ for a population generated as described before, which is the official definition of poverty line used in the EU countries. We draw a sample s_d of size $n_d = 50$, $d = 1,\ldots,D$, using sample random sampling (SRS) without replacement, independently from each area d.

A total of $L = 1,000$ population vectors $\mathbf{Y}^{(\ell)}$, $\ell = 1,\ldots,L$, were generated from the nested error model (8) with the mentioned values of model parameters and auxiliary variables. For each Monte Carlo population $\ell = 1,\ldots,L$, we calculated the true area poverty incidences and poverty gaps. Then, we selected the sample s, which is kept fixed across Monte Carlo replicates. Using the sample data $\{(Y_{di}, x_{1,di}, x_{2,di}); i \in s_d, d = 1,\ldots,D\}$ and the population data on the auxiliary variables, we computed direct estimates $\hat{F}_{\alpha d}^{\text{DIR}}$, FH, ELL, EB, Census-EB and

HB estimates of poverty incidence ($\alpha = 0$) and poverty gap ($\alpha = 1$) for each area $d = 1\ldots, D$. FH, ELL and EB estimates were obtained using REML fitting method.

For the Monte Carlo population ℓ, let $F_{\alpha d}^{(\ell)}$ be the true poverty indicator for area d and $\hat{F}_{\alpha d}^{(\ell)}$ be one of the estimates (direct, FH, ELL, EB, Census-EB or HB). Relative bias (RB) and relative root mean squared error (RRMSE) of an estimator $\hat{F}_{\alpha d}$ are approximated empirically as

$$\mathrm{RB}(\hat{F}_{\alpha d}) = \frac{L^{-1}\sum_{\ell=1}^{L}(\hat{F}_{\alpha d}^{(\ell)} - F_{\alpha d}^{(\ell)})}{L^{-1}\sum_{\ell=1}^{L} F_{\alpha d}^{(\ell)}}, \quad \mathrm{RRMSE}(\hat{F}_{\alpha d}) = \frac{\sqrt{L^{-1}\sum_{\ell=1}^{L}(\hat{F}_{\alpha d}^{(\ell)} - F_{\alpha d}^{(\ell)})^2}}{L^{-1}\sum_{\ell=1}^{L} F_{\alpha d}^{(\ell)}}.$$

For each estimator $\hat{F}_{\alpha d}$, the absolute RB (ARB) and the RRMSE are averaged across areas as

$$\overline{\mathrm{ARB}}_\alpha = D^{-1}\sum_{d=1}^{D}|\mathrm{RB}(\hat{F}_{\alpha d})|, \quad \overline{\mathrm{RRMSE}}_\alpha = D^{-1}\sum_{d=1}^{D}\mathrm{RRMSE}(\hat{F}_{\alpha d}).$$

Figure 1 depicts percent RBs (left) and RRMSEs (right) of the estimators of the domain poverty gaps F_{1d} for each area d. EB and Census-EB estimates are not shown in these plots because they are both practically equal to HB estimates and are plotted separately in Figure 2. We can see in Figure 1 left that direct, ELL and HB estimators are practically unbiased. In contrast, FH estimators display a substantial negative bias. Concerning efficiency, Figure 1 right shows that HB estimators have the smallest RRMSE whereas ELL estimators are the ones with the largest RRMSE. Conclusions for the poverty incidence F_{0d} are very similar.

Table 1 presents averages across areas of absolute RB and RRMSE of all the estimators, for both poverty incidence and poverty gap. We see that, on average, FH estimator presents a large absolute RB (over 6% for poverty incidence and close to 15% for poverty gap), whereas EB, HB and Census-HB estimators have a very small RB ($< 1\%$). The latter estimators also achieve the smallest RRMSEs (slightly over 20% for poverty incidence and over 25% for poverty gap). The largest RRMSE is obtained by ELL estimator (over 58%). Note that both absolute RB and RRMSE increase when estimating the poverty gap, because the poverty gap depends to a greater extent on the extreme of the left tail of the income distribution, which is more difficult to estimate correctly from a (finite) sample.

These results indicate that HB estimators are practically unbiased and clearly the most efficient among the considered estimators when the nested error model holds and the sample is drawn with SRS within each area. The bias of FH estimators is due to the fact that they are attaching most of the weight to the regression-synthetic component, which relies exactly on the model, but here data Y_{di} are generated from

the unit level model (8) and the area means of the covariates $\bar{X}_{k,d} = N_d^{-1} \sum_{i=1}^{N_d} x_{k,di}$ are not linearly related with the poverty indicators $F_{\alpha d}$. Thus, FH model fails due to non-linearity of the poverty indicators $F_{\alpha d}$ in the area level covariates $\bar{X}_{k,d}$, $k = 1, 2$, even if the unit level model holds exactly.

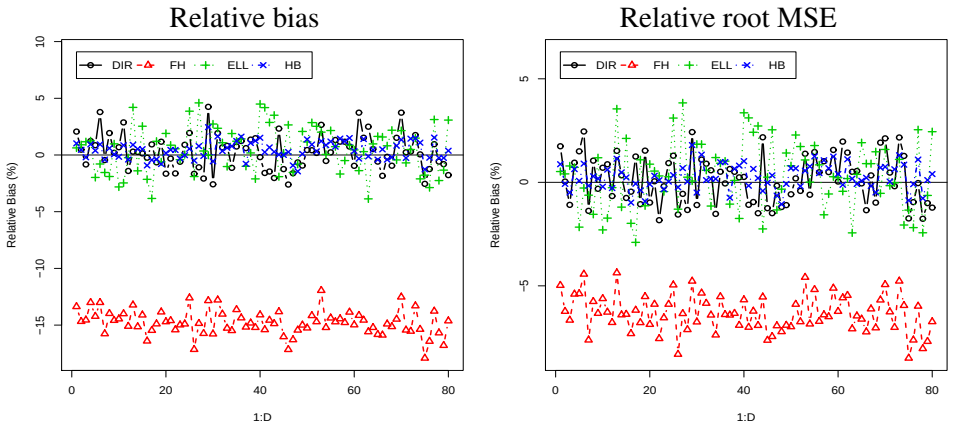


Figure 1. Percent relative bias (left) and relative root MSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1d} for each area d under the nested error model with simple random sampling.

Table 1. Averages across areas of percent ARB and RRMSE for direct, FH, HB, EB, Census-EB and ELL estimators of poverty incidence F_{0d} and poverty gap F_{1d}, under the nested error model with simple random sampling.

Method	Average ARB (%)		Average RRMSE (%)	
	F_{0d}	F_{1d}	F_{0d}	F_{1d}
Direct	0.99	1.26	28.53	36.33
FH	6.34	14.78	26.26	38.16
HB	0.48	0.65	20.15	25.43
EB	0.51	0.67	20.41	25.73
Census-EB	0.55	0.69	21.15	26.71
ELL	1.31	1.69	47.39	58.63

Figure 2 depicts percent RB (left) and RRMSE (right) of EB and Census-EB estimates of the poverty gap F_{1d} for each area d. Figure 2 left shows the great similarity of EB and Census-EB estimates of F_{1d}, even if sampling fractions in this simulation study are not so small ($n_d/N_d = 1/5$, $d = 1, \ldots, D$). See in Figure 2 right and in Table 1 that the average RRMSE increase of the Census-EB estimator is in this case less than 1%.

Next we study ELL estimator of the MSE of $\hat{F}_{\alpha d}^{\mathrm{ELL}}$. Figure 3 depicts the true MSE of ELL estimators of the poverty gap F_{1d}, labeled “True MSE ELL" and the

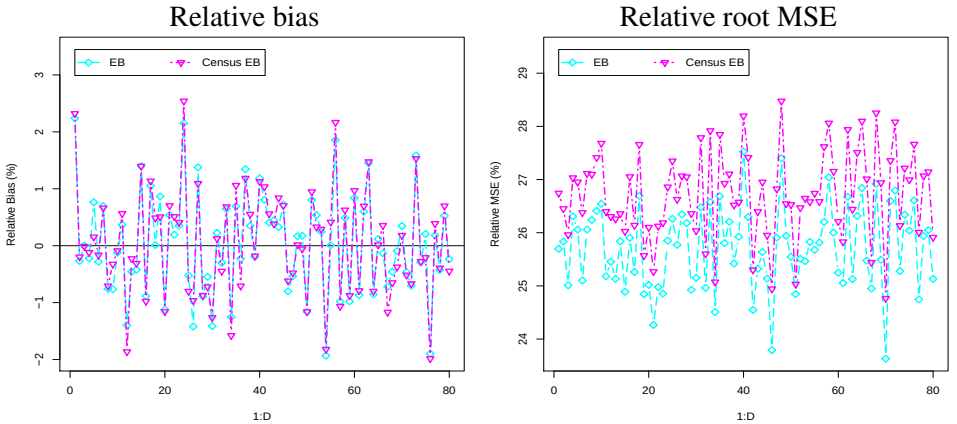


Figure 2. Percent relative bias (left) and relative root MSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1d} for each area d under the nested error model with simple random sampling.

means across simulations of ELL estimates of the MSE, labeled "MSE ELL", for each area d. This figure shows that ELL estimates of MSE do not really track the true MSEs for each area even if we have considered here random effects for the areas in the model (i.e., sampling clusters equal to areas). In the case that clusters are different from the areas, if we consider the original ELL method that includes only cluster effects but area effects are significant, then ELL estimates might seriously underestimate the MSE.

For EB estimator, the parametric bootstrap procedure proposed by Molina and Rao (2010) approximates the true MSE reasonably well, see Molina and Rao (2010). For HB estimator, posterior variance, approximated by Monte Carlo, is taken as measure of uncertainty.

4.2. Nested error model with informative sampling

We consider the same setup as in the previous simulation study, with the same population sizes, model parameters, auxiliary variables and poverty line. The only difference is that in this simulation study, samples are drawn with informative sampling. When the sampling is informative, the probability of a sample depends on the values of the population vector $\mathbf{Y}$. Thus, under this setup, the simulations need to be performed with respect to the joint distribution of $(\mathbf{Y}, s)$; that is, in each Monte Carlo replicate ℓ, we draw a population vector $\mathbf{Y}^{(\ell)}$ and, given $\mathbf{Y}^{(\ell)}$, we draw a sample $s^{(\ell)}$. A total of $L = 1000$ population vectors $\mathbf{Y}^{(\ell)}$, $\ell = 1, \ldots, L$, are generated from the true nested error model (8). Again, we consider that the target variables

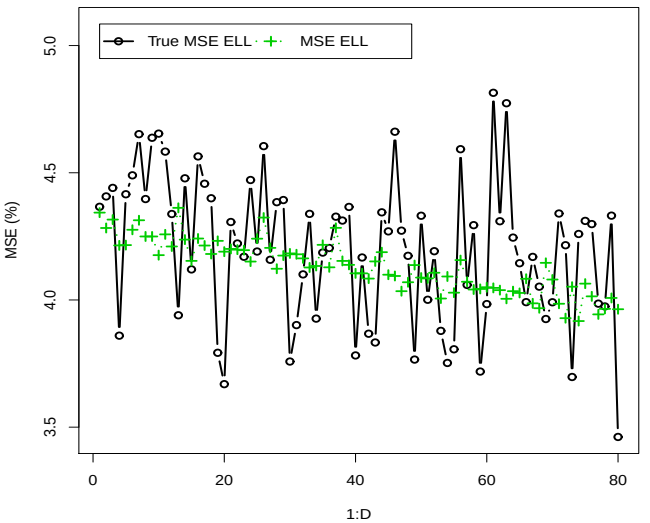


Figure 3. True MSE of ELL estimators of poverty gap F_{1d} and mean across simulations of ELL estimator of the MSE for each area d, under the nested error model with simple random sampling.

are $E_{di} = \exp(Y_{di})$. The sample $s^{(\ell)}$ is drawn by Poisson sampling, with inclusion probabilities $\pi_{d,i}$ depending on a random variable Z_{di} that is correlated with the unexplained part of Y_{di}, that is, the model errors e_{di}. Thus, for each population unit i from area d, we generate a Bernoulli random value $Q_{di} \sim \text{Bern}(\pi_{d,i})$, with $\pi_{d,i} = b^{-1}\exp(-aZ_{di})$, where $a > 0$, $b > 0$ and $Z_{di} \sim \text{Gamma}(\tau_{di}, \theta_{di})$. To choose the values of τ_{di} and θ_{di}, we consider two cases: low and high level of informativeness. In the first case, we take $\tau_{di} = 15 + 0.5e_{di}$ and $\theta_{di} = 0.75 + 0.025e_{di}$, which yield random values Z_{di} with a 20% correlation with the model errors e_{di}. In the second case, we take $\tau_{di} = 22.5 + 7.5e_{di}$ and $\theta_{di} = 1.125 + 0.375e_{di}$, yielding Z_{di} with a 80% correlation with e_{di}, which represents a high level of informativeness. Note that under informative sampling, the sample size is random because each unit in the population comes to the sample depending on its random value Q_{di}. To make this simulation study comparable with the one in previous section, we wish to have a similar average area sample size as before. This is achieved approximately by considering $a = 0.05$ and $b = 2.5$ when the informativeness level is low and taking $a = 0.02$ and $b = 4$ when the informative level is high. With the sample $s^{(\ell)}$ from each population, we compute the five estimators, namely direct, FH, EB, ELL and HB estimators. We excluded here Census-EB estimators because of their similarity with EB estimators.

Figure 4 plots RBs (left) and RRMSEs (right) of the estimators of the poverty gap F_{1d} when the informativeness level is low. Again, EB estimator is excluded because it provides nearly the same results as HB. For low level of informativeness, Figure 4 left shows that the negative bias of the EBLUP based on the FH model, observed in the simulation with SRS, still persists, while the rest of the estimators are almost unbiased. HB estimator still presents the smallest relative MSE among the considered estimators, and ELL estimator performs the worst in terms of relative MSE among the considered estimators. For the poverty incidence F_{0d}, conclusions are similar. These conclusions are confirmed by the averages across areas shown in Table 2 for both poverty incidence and poverty gap. On average, the direct estimator has the smallest absolute RB (about 0.7% for poverty incidence and 0.9% for poverty gap), followed by EB and HB estimators with a bias below 1.4% for both poverty incidence and gap, the smallest RRMSE is for EB estimator (less than 21% for poverty incidence and than 26% for poverty gap) and the largest for ELL estimator (over 47% for poverty incidence and over 58% for poverty gap).

Figure 5 plots RB (left) and RRMSE (right) of the estimators of the poverty gap F_{1d} when the level of informativeness is large. In this case, Figure 5 left shows a negative bias for the FH estimator and a large positive bias of HB and ELL estimators. Looking at Figure 5 right, we can see that now direct and FH estimates, which are calculated using the true inclusion probabilities, present the smallest RRMSE among the considered estimators. Again, conclusions are similar for the poverty incidence F_{0d}. Table 3 lists the averages across areas of ARB and RRMSE for all the considered estimators of the poverty incidence and poverty gap. In this case, the direct estimator has the smallest average ARB (about 0.6% for poverty gap), whereas the average RRMSE of ELL estimator is the largest (99.6%).

To summarize, EB and HB methods are not greatly affected under low level of informativeness, measured in terms of correlation among the design variable used in the inclusion probabilities and the response variable. When the degree of informativeness is high, these two methods are certainly affected because they do not take into account the sampling design. The effect of informative sampling on FH estimator seems to be smaller, and its negative bias is again due to a non-linearity problem of FH model because data actually follows the nested error linear regression model for log income at the unit level. We are currently developing suitable methods to handle informative sampling in the case of unit level models.

4.3. Nested error model with outliers

In this section, we carry out a simulation study under exactly the same conditions as in Section 4.1, but generating the model errors e_{di} from a mixture of normal

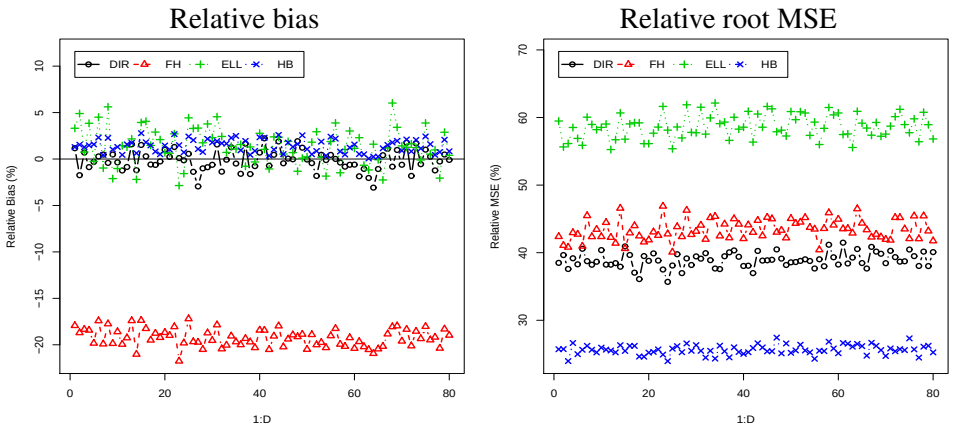


Figure 4. Percent relative bias (left) and relative root MSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1d} for each area d under low informativeness.

Table 2. Averages across areas of percent ARB and RRMSE for direct, FH, HB, EB and ELL estimators of incidence F_{0d} and poverty gap F_{1d}, under low informativeness.

Method	Average ARB (%)		Average RRMSE (%)	
	F_{0d}	F_{1d}	F_{0d}	F_{1d}
Direct	0.74	0.91	71.69	38.92
FH	10.47	19.26	30.33	43.38
HB	1.10	1.38	20.29	35.63
EB	1.04	1.25	20.48	25.86
ELL	1.63	1.98	47.39	58.65

distributions with different variances in order to create outliers. Concretely, in this simulation study, we generate model errors as $e_{di} \sim (1-\varepsilon)N(0,\sigma_e^2)+\varepsilon N(0,R\sigma_e^2)$, where ε is generated as $\varepsilon \sim \text{Bern}(p)$. We consider two fractions of outliers, $p=0.1$ and $p=0.5$, and two values for the factor R in the variance of outliers, namely $R=10$ and $R=100$. Using the above mechanism to generate model errors, a total of $L=1000$ population vectors $\mathbf{Y}^{(\ell)}, i=1,\ldots,L$, were generated from the nested error model (8). Then, we calculated true area poverty incidences and gaps. Note that the outliers considered in this simulation study are not recording errors in the sample data. They are actually representative outliers appearing in the population. Thus, they are actual realizations of the distribution with heavier tails obtained from the normal mixture, and true values of poverty indicators actually include the generated outliers in the population. The sample is drawn by SRS within each area as in Section 4.1, keeping the sample units s fixed across simulations. With each Monte Carlo sample, direct, FH, EB, ELL and HB estimators were computed.

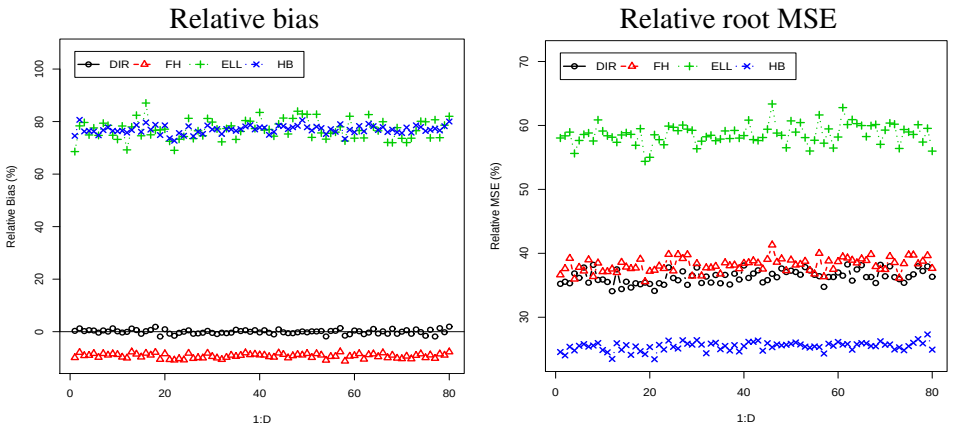


Figure 5. Percent relative bias (left) and relative root MSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1d} for each area d under high informativeness.

Table 3. Averages across areas of percent ARB and RRMSE for direct, FH, HB, EB and ELL estimators of incidence F_{0d} and poverty gap F_{1d}, under large informativeness.

Method	Average ARB (%)		Average RRMSE (%)	
	F_{0d}	F_{1d}	F_{0d}	F_{1d}
Direct	0.59	0.65	23.62	25.69
FH	6.94	9.21	23.83	29.40
HB	61.64	76.95	66.05	84.95
EB	61.60	73.68	66.08	84.89
ELL	61.69	76.98	72.94	97.29

We report here results for the cases of less frequent mild outliers ($p = 0.1$ and $R = 10$), and of more frequent and extreme outliers ($p = 0.5$ and $R = 100$). For the first case, results for the poverty gap are plotted in Figure 6. Again, EB is excluded in the plots because it provides similar results as HB. Figure 6 left and right show that direct estimators are not practically affected by the outliers, which is expected because this estimator does not rely on any model assumption. Similarly, FH estimator is less affected by outliers because the observed negative bias is again due to non-linearity problems. HB and ELL estimators show a moderate bias, but still HB estimator achieves the lowest error in terms of RRMSE. Averages across areas of ARB and RRMSE for all estimators of poverty incidence and poverty gap are shown in Table 4. We can see that the bias of EB and HB estimators is small (around 4% for poverty incidence and 5% for poverty gap), and the RRMSE has increased only about 0.5% with respect to the case of no outliers (see Table 2) and it is still acceptable (around 21% for poverty incidence and 26% for poverty gap).

For the case of more frequent and extreme outliers ($p = 0.5$ and $R = 100$), Figure 7 left shows that in this case HB, and to a greater extent ELL estimators; present a very large positive bias, see also Table 7 reporting averages across areas. Note that the RRMSE of ELL estimator reaches 226.63% for the poverty gap. In this simulation study, FH estimates perform better than in the previous simulation studies, and this could be due to the fact that, since FH model is less correct when outliers are present, the FH estimator is attaching more weight to the direct estimator, which is practically unbiased. EB, HB and ELL estimators are severely biased when data contains frequent extreme outliers, performing even worse than under high level of informative sampling, but are not too much affected under rare and not so extreme outliers. These methods are based on model assumptions and are not robust to strong model misspecification when the true error distribution has very heavy tails as in the mixture model considered here with $p = 0.5$ and $R = 100$. We are exploring estimation methods for complex parameters that are robust to outliers. Note that previous work on robust estimation, e.g. Sinha and Rao (2009), focused on estimating area means only.

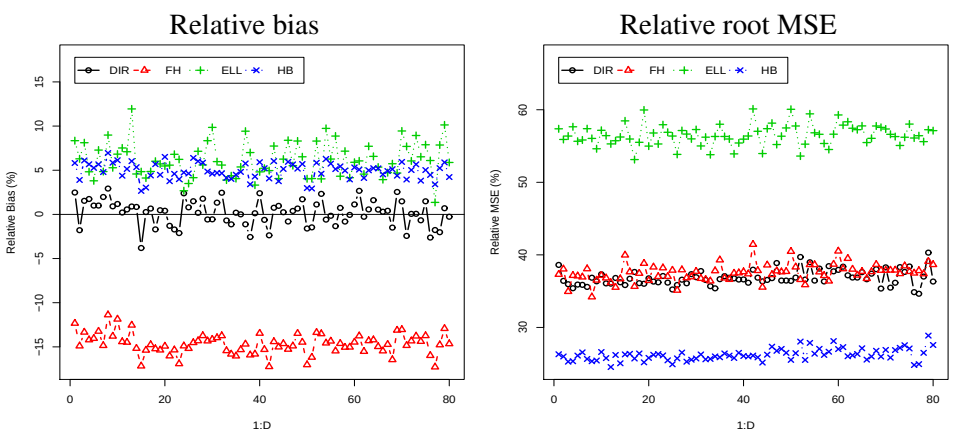


Figure 6. Percent relative bias (left) and relative root MSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1d} for each area d under nested error model with outliers ($p = 0.01$ and $R = 10$).

5. Conclusions

This paper reviews popular poverty mapping procedures focusing on practical aspects. Simulation studies compare these methods under three interesting scenarios that show the good properties when assumptions hold and also the worse performance when some assumptions are not satisfied. These simulation studies illustrate

Table 4. Averages across areas of percent ARB and RRMSE for direct, FH, HB, EB and ELL estimators of incidence F_{0d} and poverty gap F_{1d}, under under nested error model with outliers ($p = 0.01$ and $R = 10$).

Method	Average ARB (%)		Average RRMSE (%)	
	F_{0d}	F_{1d}	F_{0d}	F_{1d}
Direct	0.92	1.18	28.54	36.82
FH	6.16	14.67	26.10	37.55
HB	3.95	4.95	20.81	26.22
EB	3.88	4.79	20.99	26.42
ELL	4.93	6.14	46.65	56.52

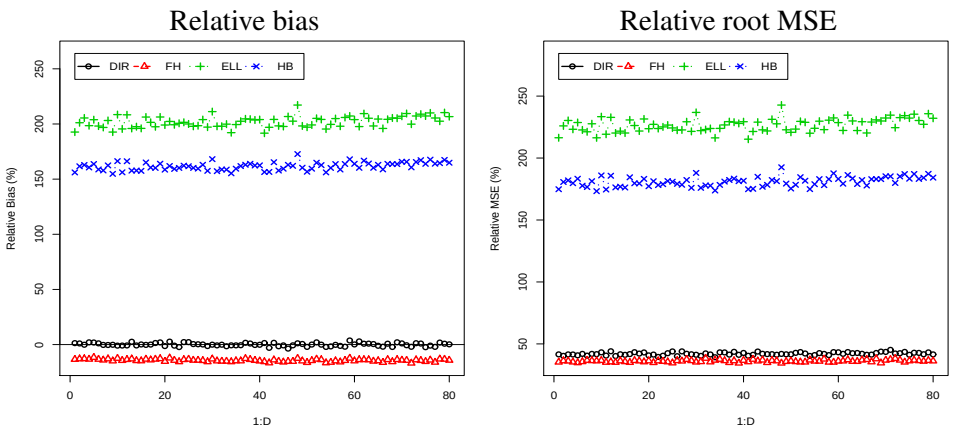


Figure 7. Percent relative bias (left) and relative root MSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1d} for each area d under nested error model with outliers ($p = 0.05$ and $R = 100$)

that: (i) Even if aggregation protects against model failures in FH area level model, the linearity assumption of the model fails when data follows a unit level model but target parameters are nonlinear functions of the model responses. However, FH estimates are less affected by informative sampling and by symmetric representative unit level outliers. (ii) EB and HB methods perform practically the same, and are the best among the considered estimators when the nested error model with normality holds and sampling is noninformative. They are not very much affected by mildly informative sampling and small proportion of mild outliers, but might be severely affected by highly informative sampling or severe outliers in large proportions. (iii) Census-EB estimators of poverty indicators are practically the same as EB estimators and avoid linking the survey and census data files. (iv) ELL method under a nested error model with random area effects performs the worst in all scenarios

Table 5. Averages across areas of percent ARB and RRMSE for direct, FH, HB, EB and ELL estimators of incidence F_{0d} and poverty gap F_{1d}, under nested error model with outliers ($p = 0.05$ and $R = 100$).

Method	Average ARB (%)		Average RRMSE (%)	
	F_{0d}	F_{1d}	F_{0d}	F_{1d}
Direct	0.96	1.20	29.68	41.99
FH	5.66	14.33	26.65	36.10
HB	74.13	161.73	86.87	180.88
EB	74.11	161.59	86.95	180.81
ELL	92.64	201.97	111.32	226.63

because it does not account for unexplained between-area variation.

Several relaxations of the normality assumption in the EB method have been recently studied. Diallo and Rao (2004) derived EB estimators of poverty indicators assuming the family of skew normal (SN) distributions for the random effects and/or the errors, which includes the normal distribution as a particular case. Their results indicate that the EB method based on normality is robust to deviations from normality of u_d provided e_{di} remains normal. On the other hand, under SN errors e_{di}, normality-based EB estimators can induce significant bias and may not perform well compared to SN-based EB estimators. Van der Weide and Elbers (2014) studied normal mixture models on the area effects u_d and the errors e_{di}. Their results are in agreement with Diallo and Rao (2014) in the sense that the normality-based EB method is robust provided e_{di} remains normal. Graf, Marín and Molina (2015) have also extended the EB method to the generalized Beta distribution of the second kind (GB2), which models income data adequately. They have also shown that using the EB method based on the GB2 distribution leads to clear efficiency gains when the distribution of log income deviates from normality, whereas it does not lose efficiency when log incomes follow the nested error model with normality.

Acknowledgements

Isabel Molina was supported by grants ref. MTM2009-09473, MTM2012-37077-C02-01 and SEJ2007-64500 and J.N.K. Rao's research by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

BATES, D., MAECHLER M., BOLKER, B., WALKER, S., (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7.

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

CORREA, L., MOLINA, I., RAO, J.N.K., (2012). Comparison of methods for estimation of poverty indicators in small areas. Unpublished report.

DIALLO, M., RAO, J.N.K., (2014). Small Area Estimation of Complex Parameters Under Unit-level Models with Skew-Normal Errors. Proceedings of the Survey Research Section, American Statistical Association.

ELBERS, C., LANJOUW, J. O., LANJOUW, P., (2003). Micro-level Estimation of Poverty and Inequality. *Econometrica*, 71(1), 355–364.

FAY, R., HERRIOT R., (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of American Statistical Association,* 74, 269–277.

FOSTER, J., GREER, J., THORBECKE, E., (1984). A class of decomposable poverty measures. *Econometrica*, 52, 761–766.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D., and SANTAMARÍA, L., (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78, 443–462.

GRAFF, M., MARÍN, J.M., MOLINA, I., (2015). Estimation of poverty indicators in small areas under skewed distributions, Unpublished manuscript.

MOLINA, I., MORALES, D., (2009). Small area estimation of poverty indicators. *Boletín de Estadística e Investigación Operativa*, 25, 318–325.

MOLINA, I., MARHUENDA, Y., (2015), Sae: An R Package for Small Area Estimation, *R Journal*, in print.

MOLINA, I., RAO, J.N.K., (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38, 369–385.

MOLINA, I. NANDRAM, B. and RAO, J.N.K., (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. *The Annals of Applied Statistics*, 8(2), 852–885.

PFEFFERMANN, D., (2013). New important developments on small area estimation. *Statistical Science*, 28, 40–68.

RAO, J.N.K., (2003). *Small Area Estimation*. Hoboken, NJ: Wiley.

RAO, J.N.K., MOLINA, I., (2015). *Small Area Estimation, Second Edition*. Hoboken, NJ: Wiley, in print.

SINHA, S., RAO, J.N.K., (2009). Robust small area estimation. *The Canadian Journal of Statistics*, 37, 381–399.

VAN der WEIDE, R., ELBERS, C. (2013). Estimation of normal mixtures in a nested error model with an application to small area estimation of welfare. Speech presented at the SAE Conference 2013, Bangkok, Thailand.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 67–90

A TWO-COMPONENT NORMAL MIXTURE ALTERNATIVE TO THE FAY-HERRIOT MODEL

Adrijo Chakraborty[1], **Gauri Sankar Datta**[2 3 4], **Abhyuday Mandal**[5]

ABSTRACT

This article considers a robust hierarchical Bayesian approach to deal with random effects of small area means when some of these effects assume extreme values, resulting in outliers. In the presence of outliers, the standard Fay-Herriot model, used for modeling area-level data, under normality assumptions of random effects may overestimate the random effects variance, thus providing less than ideal shrinkage towards the synthetic regression predictions and inhibiting the borrowing of information. Even a small number of substantive outliers of random effects results in a large estimate of the random effects variance in the Fay-Herriot model, thereby achieving little shrinkage to the synthetic part of the model or little reduction in the posterior variance associated with the regular Bayes estimator for any of the small areas. While the scale mixture of normal distributions with a known mixing distribution for the random effects has been found to be effective in the presence of outliers, the solution depends on the mixing distribution. As a possible alternative solution to the problem, a two-component normal mixture model has been proposed, based on non-informative priors on the model variance parameters, regression coefficients and the mixing probability. Data analysis and simulation studies based on real, simulated and synthetic data show an advantage of the proposed method over the standard Bayesian Fay-Herriot solution derived under normality of random effects.

Key words: Hierarchical Bayes, heavy-tail distribution, non-informative priors, robustness to outliers; small area estimation.

[1]NORC at the University of Chicago, Bethesda, MD 20814. E-mail:chakraborty-adrijo@norc.org
[2]Department of Statistics, University of Georgia, Athens, GA 30602, USA.
E-mail:gauri@stat.uga.edu
[3]Center for Statistical Research and Methodology, US Census Bureau, Washington, D.C. 20233
[4]Disclaimer: This report is released to inform interested parties of research and to encourage discussion on work in progress. The views expressed are those of the authors and not necessarily those of the US Census Bureau
[5]Department of Statistics, University of Georgia, Athens, GA 30602, USA.
E-mail:amandal@stat.uga.edu

1. Introduction

Small area estimation methods are becoming increasingly popular among survey practitioners. Reliable small area estimates are often solicited by policy makers from both government and private sectors for planning, marketing and decision making. In order to meet the growing demand for reliable small area estimates, researchers have developed methods that combine information from small areas and other related variables. Ghosh and Rao (1994), Rao (2003), Jiang and Lahiri (2006), Datta (2009) and Pfeffermann (2013) provided a comprehensive review of the research in small area estimation.

The landmark paper by Fay and Herriot (1979) used the empirical Bayes (EB) approach (see, for example, Efron and Morris, 1973) and popularized model-based small area estimation methods. Denoting the design-based direct survey estimator of the ith small area by Y_i and its auxiliary variable by x_i, an $r \times 1$ vector, Fay and Herriot (1979) introduced the model

$$Y_i = \theta_i + e_i, \quad \theta_i = x_i^T \beta + v_i, \quad i = 1, \ldots, m. \tag{1.1}$$

Here θ_i is a summary measure of the characteristic to be estimated for the ith small area, e_i is the sampling error of the estimator Y_i, and the random effects v_i denote the model error measuring the departure of θ_i from its linear regression on x_i. It is assumed that $e_1, \ldots, e_m$ are independent and normally distributed with $e_i \sim N(0, D_i)$, and are independent of $v_1, \ldots, v_m$, which are i.i.d. $N(0, A)$. The sampling variances D_i's are treated as known, but the model parameters β and A are unknown. Random effects v_i's are also known as small area effects.

In this paper we focus on hierarchical Bayes (HB) methods for area-level models. The classical area-level Fay-Herriot model was primarily developed as a frequentist model, which was later given a Bayesian formulation (Rao 2003; Datta et al. 2005). Estimators obtained from the Fay-Herriot model are shrinkage estimators, i.e., a weighted average of the direct estimator and the model-based synthetic estimator, and these weights depend on the model assumption. Datta and Ghosh (2012) gave an extensive review of shrinkage estimation in the small area estimation context. Shrinkage estimators are primarily constructed to improve standard estimators. For instance, in the small area context model based shrinkage estimators are constructed

to improve the precision of direct estimators such as the sample mean or the Horvitz-Thompson estimator. Datta and Lahiri (1995) discussed how outliers can affect shrinkage estimators, claiming that even a single outlier may lead all the small area estimates to collapse to their corresponding direct estimates. This phenomenon was also mentioned in the context of estimation of multiple normal means under the assumption of an exchangeable normal prior (cf. Efron and Morris 1971, Stein 1981, and Angers and Berger 1991). One or more substantive outliers considerably inflate the standard estimator of model variance.

An overestimation of model variance due to one or more substantive outliers practically results in no shrinkage of any of the direct estimates of the small area means to the synthetic regression estimator. This also limits the reduction in the posterior variances of the model-based estimates. To rectify this problem, following the work of Angers and Berger (1991), who used a Cauchy distribution for the small area means θ_i, Datta and Lahiri (1995) recommended a broader class of heavy-tailed distributions through a scale mixture of normal distributions. They showed that under these assumptions, in the presence of substantive outliers, estimators corresponding to the outlying areas converge to their corresponding direct estimators but leave the non-outlying areas less affected. One difficulty with the last method is that the mixing distribution for the scale parameter is considered to be known. For example, one can use t-distribution for random effects, as in Xie et al. (2007). However, in the absence of any information regarding the degrees of freedom, one needs to specify a prior. Xie et al. (2007) assumed a gamma prior for the degrees of freedom. The hyperparameters involved in this gamma distribution need to be specified. Bell and Huang (2006) argued that, under practical circumstances, limited information is obtained from the data regarding the degrees of freedom, and instead they used several fixed values for the degrees of freedom.

In order to avoid specifying the mixing distribution in the previous paragraph, in this paper we propose a two-component normal mixture distribution for the random small area effects. Our model accommodates means for outlying areas to come from the distribution with a larger variance. This is a simple extension of the Fay-Herriot model with a contaminated random effects distribution with possibly small proportion of areas having a larger model variance. Contaminated models have been extensively used in empirical evaluations of the robust empirical best linear

unbiased prediction (EBLUP) approach of Sinha and Rao (2009). We consider an HB approach by assigning non-subjective priors to the parameters involved in the model. Some components of these priors are improper, hence we provide sufficient conditions for the posterior distribution to be proper.

In a recent article, Datta et al. (2011) demonstrated that in the presence of good covariates x_i, the variability of the small area means θ_i may be accounted for well by x_i, and including a random effects v_i in the model (1.1) may be unnecessary. These authors test a null hypothesis of no random effects in the small area model and if it is not rejected, they propose more accurate synthetic estimators for the small area means. In a more recent article, Datta and Mandal (2015) argued that even if the null hypothesis was rejected in this case, it would be reasonable to expect only a small fraction of the small areas means would not be adequately explained by the covariates, and only these areas would require a random component to the regression model.

Using the HB approach, Datta and Mandal (2015) considered a "spike and slab" distribution for the random small area effects in order to propose a flexible balance between the Fay and Herriot (1979) and Datta et al. (2011) models. However, it is often difficult to find reliable covariates that would describe the response well, particularly, if the number of small areas is large. For such datasets, not only the test proposed by Datta et al. (2011) would suggest the inclusion of the small area effects, but also the model proposed by Datta and Mandal (2015) would estimate the probability of the existence of random effects as very high. This would effectively suggest the Fay-Herriot model, but, in reality, only a small proportion of small areas may not be adequately explained by a model with one single A. This would result in an overestimation of A, thereby resulting in a poor fit, particularly when the number of small areas m is large. Even if most of the small areas would require a random effects term in the regression model, it is more likely that only a small proportion of small areas would need a bigger value of A, and a smaller value of the same would be sufficient for other areas. In this paper, we assume that $v_1, \ldots, v_m$ are independently distributed with mean 0 and a two-component mixture of normal distributions with variance either A_1 or $A_2 (> A_1)$. This model is potentially useful for handling large outliers in small area means.

Bell and Huang (2006) presented an insightful discussion about using a t-distribution

with a known d.f. to handle outliers in the Fay-Herriot model. The theoretical regression residuals from (1.1) consist of the sum of the sampling error and the model error, which are not individually observable. Bell and Huang (2006) argued that a residual may be an outlier, either due to the sampling error or the model error. It is difficult to distinguish between the scenarios of the sampling error outlier or the model error outlier, since the data used in fitting the model (1.1) cannot readily disentangle the two cases. They explained that the consequences of these two types of outliers are quite different. If the model error v_i is an outlier for some areas, then the regression model (or synthetic estimation) is not good for these areas. In that case, the direct estimator Y_i should be used as the small area estimator. Datta and Lahiri (1995) considered this case using a scale mixture of normal distribution. An alternative to this approach is proposed in the present article through a two-component normal mixture. Bell and Huang (2006) noted that, in the presence of a model outlier, if the direct estimator also has large variability, then no satisfactory solution exists. On the other hand, if the sampling error e_i is an outlier due to an underestimation of the variance D_i, then the direct estimator Y_i is not reliable; Bell and Huang (2006) argued that the "synthetic estimator" $x_i^T\beta$ may be used for prediction. To address this issue, they proposed a t-distribution for the sampling distribution. For further discussion, we refer to this article.

There is a substantive literature on the frequentist approach for the robust estimation of small area means in the presence of outliers. Ghosh et al. (2008) considered the robust empirical Bayes estimation of small area means for area level model. They used the Huber's ψ-function to limit the influence of outliers. For unit level models Sinha and Rao (2009) and Chambers et al. (2014) proposed a robust modification of EBLUPs of the finite population means of small areas. They also used the Huber's ψ-function to limit the impact of outlier observations on the estimators of model parameters and the best linear unbiased predictors. While Sinha and Rao (2009) provided robust projective EBLUPs (in the terminology of Chambers et al. (2014)) of the finite population small area means, the latter group of authors discussed the limitation of such predictors in terms of bias, and also proposed robust predictive EBLUPs to remedy this concern.

This paper is organized as follows. In Section 2 we describe the proposed model and discuss some properties of our new shrinkage estimators. In Section 3 we illus-

trate our method to estimate U.S. poverty rates for 3141 counties, based on 5-year estimates from the American Community Survey. The performance of the model, in comparison with the traditional Fay-Herriot model, is discussed in Section 4 and Section 5. Section 6 provides a concluding discussion. A detailed proof of the propriety of the posterior distribution is moved to the Appendix.

2. Two-component normal mixture model

Fay and Herriot (1979) proposed a model which has been extensively used in many small area estimation applications to provide reliable estimates of poverty and income measures. While for regular data the model successfully produces accurate shrinkage estimators of small area means, it breaks down in the presence of substantial outliers among small area means. In order to account for the outliers, we consider a two-component normal mixture extension of the Fay-Herriot model. This model is given by

$$y_i = \theta_i + e_i, \;\; \theta_i = x_i^T \beta + (1-\delta_i) v_{1i} + \delta_i v_{2i}, \; i = 1, \ldots, m, \tag{2.1}$$

where e_i, δ_i, v_{1i}, v_{2i} are independently distributed with $P(\delta_i = 1|p) = 1-p$, $v_{1i} \sim N(0, A_1)$ and $v_{2i} \sim N(0, A_2)$. As in (1.1), β is an $r \times 1$ vector of regression parameters, and the sampling errors $e_1, \ldots, e_m$ are independently normally distributed. To complete our HB structure, we consider the following class of priors,

$$\pi(\beta, A_1, A_2, p) = \pi^*(A_1, A_2) \propto A_1^{-\alpha_1} A_2^{-\alpha_2} I(0 < A_1 < A_2 < \infty). \tag{2.2}$$

We use a uniform prior on the regression parameter β and the mixing proportion p. For the prior on the variance parameters, we choose $\alpha_1 < 1 < \alpha_2$ suitably, and we discuss the permissible choices of the values of α_1 and α_2 later. We impose the restriction $A_1 < A_2$, so that we do not have a label switching problem leading to non-identifiability. The area-specific random effects corresponding to the outlying areas in the model are assumed to follow a normal distribution with larger variance, which remains the motivation behind imposing such a restriction. While for the parameter β common to all the components of the mixture model, an improper uniform prior is reasonable, the prior for A_1 and A_2, which are not common in all the components of the mixing distributions, is required to be at least *partially proper*. By partially

proper we mean that while the marginals are improper, conditional priors for A_2 given A_1, and A_1 given A_2 are proper. For this to hold for our class of priors for A_1,A_2, it is necessary and sufficient that $\alpha_1 < 1 < \alpha_2$. A partially proper prior is required for the parameters that are not common to all components of a Bayesian mixture model (cf. Scott and Berger, 2006).

Since the Bayesian model involves improper priors, in Theorem 2.1 below we provide sufficient conditions that ensure the resulting posterior distribution from the proposed model will be proper. A detailed proof of Theorem 2.1 is given in Section 6.

Theorem 2.1 *The resulting posterior distribution from model (2.1) and the prior in (2.2) will be proper if (a)* $m > r + 2(2 - \alpha_1 - \alpha_2)$ *and (b)* $2 - \alpha_1 - \alpha_2 > 0$.

The sufficient conditions in Theorem 2.1 provide a set of permissible values for α_1 and α_2. In conjunction with the condition $2 - \alpha_1 - \alpha_2 > 0$, the condition $\alpha_2 > 1$ implies $\alpha_1 < 1$. We noted earlier that the last two conditions are necessary to elicit partially proper priors. The special case $\alpha_1 = 0$ is feasible, which corresponds to a uniform prior, provided $1 < \alpha_2 < 2$. However, it is not possible to assign a uniform prior on A_2. If $\alpha_1 = \frac{1}{2}$, then $1 < \alpha_2 < \frac{3}{2}$. Also, for mixture models, Jeffreys' prior has no closed-form expression to work with.

Our choice of a prior for the mixing parameter p is Uniform(0,1). We can modify this prior if subjective information is available. If past experience in an application suggests any information regarding the proportion of the outlying areas, it can be incorporated in the model by modifying the prior for p. Sufficient conditions for the propriety of the posterior density will remain unchanged. For instance, if the model is modified with the assumption that p follows a known *Beta* distribution, the sufficient conditions provided in Theorem 2.1 will remain intact.

It is well-known that even a single substantial outlier will collapse shrinkage estimators of all θ_i's based on the model (1.1) to the direct estimators y_i's (see Dey and Berger, 1983; Stein, 1981). As a result, model-based estimators will fail to borrow strength from other small areas. To protect against this odd behaviour, Angers and Berger (1991), and Datta and Lahiri (1995) suggested a robust shrinkage

model. These authors used a suitable scale mixture of normal distributions to model a long-tail distribution of the θ's. These methods assume the knowledge of the scale mixing distribution, which may not be available. The purpose of our mixture model proposed in (2.1) is to provide an alternative solution that does not require the knowledge of the mixing distribution and to facilitate borrowing information among non-outlying observations in the presence of some substantive outliers.

Below we discuss a heuristic comparison of the shrinkage property of the Bayes estimators of θ_i under the Fay-Herriot model and our proposed model, in the presence of substantial outliers. For the Fay-Herriot model, given the values of the parameters β and A, an estimator of θ_i is

$$\theta_i^{FH} = y_i - \frac{D_i}{D_i+A}(y_i - x_i^T\beta),\ i=1,\ldots,m. \tag{2.3}$$

In the presence of outliers, the frequentist estimators of A will be large, and the posterior density of A will have a long right tail, which will also result in a large Bayesian estimator of A. Consequently, an estimate of the shrinkage coefficient $D_i/(D_i+A)$ will be rather small, and the Bayes or the EB estimator of θ_i will borrow little from its synthetic regression prediction and it will collapse to direct estimator y_i for all i.

We now argue that the proposed mixture model is more flexible to retain shrinkage of the non-outlying observations in the presence of outliers. Let $E(\theta_i|\beta,A_1,A_2,p,y) = \theta_i^{Mix}$. Using iterated expectation $E(\theta_i|\beta,A_1,A_2,p,y) = E[E(\theta_i|\beta,A_1,A_2,\delta_i,p,y)|\beta,A_1,A_2,p,y]$, and after noting that $E(\theta_i|\beta,A_1,A_2,\delta_i,p,y) = \dfrac{D_i x_i^T\beta + A_{1+\delta_i}y_i}{D_i+A_{1+\delta_i}}$, $\tilde{p}_i = P(\delta_i=0|\beta,A_1,A_2,p,y)$, we get

$$\theta_i^{Mix} = y_i - \left[\left(\frac{D_i}{D_i+A_1}\right)\tilde{p}_i + \left(\frac{D_i}{D_i+A_2}\right)(1-\tilde{p}_i)\right](y_i - x_i^T\beta), \tag{2.4}$$

where

$$\tilde{p}_i = \frac{\frac{p}{(D_i+A_1)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\frac{(y_i-x_i^T\beta)^2}{(D_i+A_1)}\right\}}{\frac{p}{(D_i+A_1)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\frac{(y_i-x_i^T\beta)^2}{(D_i+A_1)}\right\} + \frac{(1-p)}{(D_i+A_2)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\frac{(y_i-x_i^T\beta)^2}{(D_i+A_2)}\right\}}, \tag{2.5}$$

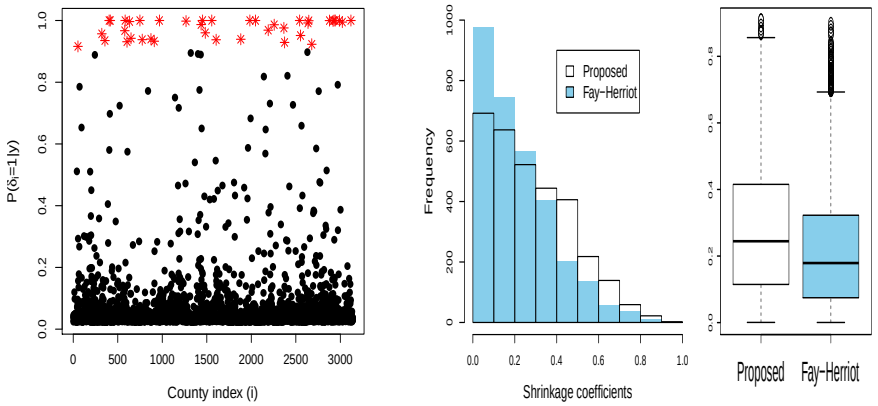
for $i=1,\ldots,m$. In the presence of substantially large outliers, $(y_i - x_i^T\beta)^2$ and A_2

are expected to be high, hence $P(\delta_i = 0|\beta, A_1, A_2, p, y_i) \approx 0$. This will result in the second shrinkage term within square brackets in (2.4) to be dominant. However, since the posterior distribution of A_2 has a long tail, the shrinkage coefficient associated with the second component will be small and $\theta_i^{Mix} \approx y_i$, i.e., if the i^{th} area is outlying then the small area estimator based on this model will be very close to its direct estimator. On the other hand, for any non-outlying areas $\tilde{p}_i$ will be away from 0, and their shrinkages will be less impacted by the outliers.

3. Data Analysis

We illustrate our proposed methodology by analysing a real data obtained from the "American Fact Finder" website maintained by the US Census Bureau. The data set contains 5-year ACS estimates of the overall poverty rates for 3141 US counties along with their associated design-based standard errors. The county identifiers are not available for confidentiality reasons. In order to improve direct design-based estimates, government agencies implement state-of-the-art small area estimation methods to produce model-based estimates using auxiliary data. For poverty estimation, the domain-level tax data are typically used as auxiliary information. However, tax data are not available for public use, due to legal restrictions. In our analysis we use the foodstamp participation rate as our only auxiliary variable (the correlation between the foodstamp participation rate and the overall poverty rate is 0.81). Initially we fit the Fay-Herriot model (1.1) with the restricted maximum likelihood method (REML) as well as the hierarchical Bayesian (HB) method, assuming flat priors for regression and variance parameter. The REML and Bayes estimates of the model parameters are very close: $\hat{\beta}^{REML} = (0.056, 0.634)^T$, $\hat{A}^{REML} = 0.0009$ and $\hat{\beta}^{Bayes} = (0.051, 0.634)^T$, $\hat{A}^{Bayes} = 0.0009$.

We have applied the proposed method to this data set and report the results in Table 1. Our choices of α_1 and α_2 are 0.3 and 1.3 respectively. We have also performed further analysis with other choices of α_1 and α_2 within the feasible range, but the results were not considerably different. From Table 1, we see that the posterior mean of A_2(= 0.00619) is almost ten times larger than that of A_1(= 0.00054). In addition, the estimate $\hat{p} = 0.07$ indicates that there are about 7% of small areas which have much larger area specific variability compared to the majority. The outlying



Bayes estimates of $P(\delta_i = 1|y)$ Shrinkage coefficients

Figure 1: Analysis of the American Community Survey data

areas can be identified by computing the Bayes estimates of posterior probabilities $P(\delta_i = 1|y)$. We plot the estimates of these probabilities for each area in Figure 1. It shows that although most areas have low probabilities of having high random effects, some of them have higher chances of having a large variability in the model error or the random small area effects. According to our analysis, approximately 7% (221 out of 3141) of small areas have the posterior probability $P(\delta_i = 1|y) > 0.15$, and approximately 1.3% (40 out of 3141) of small areas have the posterior probability $P(\delta_i = 1|y) > 0.9$.

4. Exploration of the shrinkage coefficients

We compare the shrinkage coefficients resulting from the proposed method with those resulting from the standard Fay-Herriot model. By simulations we demonstrate that the proposed method usually provides better shrinkage than the Fay-Herriot method in the presence of outliers in the data. On the other hand, simulated data from the standard Fay-Herriot model yield shrinkage coefficients based on the proposed model that are very similar to those based on the Fay-Herriot model.

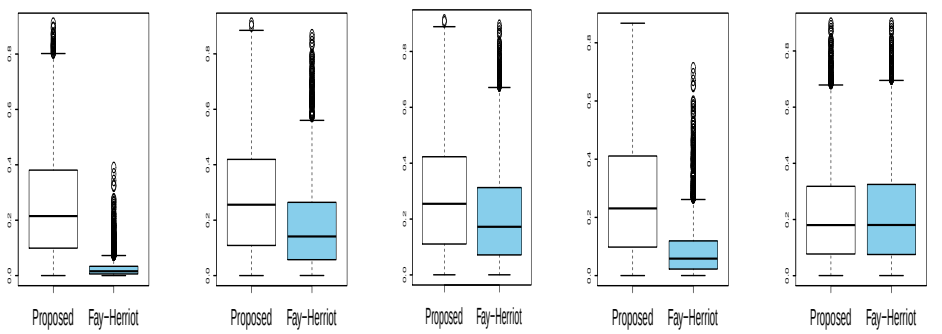
Table 1: HB estimates of model parameters (for the ACS county level poverty rates data)

	Posterior	Posterior	Posterior Quantiles		
Parameter	Mean	sd	2.5%	Median	97.5%
β_1	0.0465	0.0013	0.0440	0.0465	0.0491
β_2	0.6605	0.0075	0.6459	0.6607	0.6748
A_1	0.00054	0.00003	0.00049	0.00054	0.00059
A_2	0.00619	0.00103	0.00454	0.00609	0.00854
p	0.0725	0.0237	0.0470	0.0704	0.1037

These two simulations, presented in Figure 2 essentially show the robustness of the proposed method to outliers.

We mentioned in Section 2 that the proposed method is expected to provide better overall shrinkage than Fay-Herriot method in the presence of outliers. In order to demonstrate this property of the model, we conduct the following simulations. We replace the direct estimates of the first 10% of small areas of the data by simulated values and retain the rest of the data set intact. The purpose is to artificially contaminate the data set. We generate the direct estimates of the first 10% of small areas from the model (1.1). We use the sampling variances of these areas to generate the corresponding sampling errors. We use the estimated regression parameters $\beta = (0.06, 0.6)^T$ and model variance 0.0009 obtained from the Fay-Herriot analysis of the original data, using the Prasad-Rao method. We use these model parameter values and the values of the auxiliary variables from these 10% of small areas to retain the mean structure and variability of the small area means which are nearly similar to the original population. We introduce outliers through the use of a heavy tail distribution or large model variance for random effects. Random small area effects are generated from (a) $v_i \sim t_1$, (b) $v_i \sim t_2$, (c) $v_i \sim t_3$, with proper scaling for each and (d) $v_i \sim N(0, 5^2 \times a^2)$. Note that t_1 distribution is the Cauchy distribution which does not have a variance (indeed it does not have a mean either). We rescale the draws from t_1, t_2 and t_3, multiplying them by the adjusting factor, $\frac{N_{0.75}}{T^{df}_{0.75}} a$, where $N_{0.75}$ and $T^{df}_{0.75}$ are the 75^{th} percentile of $N(0, 1^2)$ and t (for a specified df) respectively. By multiplying the draws by this adjusting factor, we intend to match the

inter-quartile range of draws from the t-distribution to the inter-quartile range of a $N(0, a^2)$ distribution. Since the Prasad-Rao estimate of the random effects variance based on the original data is 0.0009, we choose $a^2 = 0.0009$ in order to maintain consistency.



(a) (b) (c) (d) (e)

Figure 2: *Boxplots of the estimated shrinkage coefficients for two methods. In plots (a)−(d), data are partially simulated for some small areas by drawing random effects from (a) t_1, (b) t_2, (c) t_3, (each of (a)−(c) scale adjusted) and (d) $N(0, 5^2 \times (0.03)^2)$. In plot (e), we fully simulate data for all areas by drawing random effects from $N(0, (0.03)^2)$.*

We apply the proposed method, as well as the Fay-Herriot method, and compare the estimates of shrinkage coefficients in Figures 2 and 3. We see from Figure 3 that when we partially contaminate the data set using (a) re-scaled t_1 (Cauchy) and (d) $\mathrm{N}(0, 5^2 \times (0.03)^2)$, the overall shrinkage obtained from the proposed model is considerably higher than the overall shrinkage obtained from the regular Fay-Herriot method. This result shows the flexibility of the proposed model in borrowing information from other areas when outliers in the random effects are present. Panels (b), (c) and (e) of Figure 2 show that the proposed method performs similarly to the Fay-Herriot method when the departure of the random effects distribution from the normal is moderate or none.

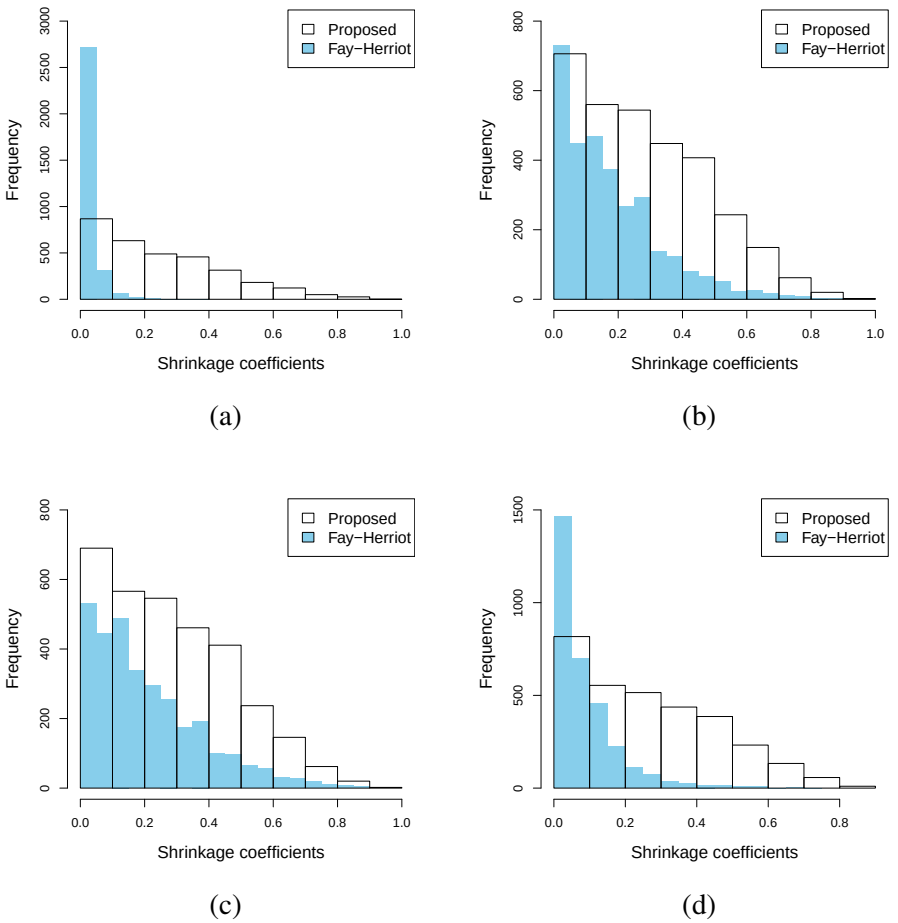


Figure 3: *Histograms of the estimated shrinkage coefficients of the two methods when the data are partially simulated by drawing random effects from (a)* t_1*, (b)* t_2*, (c)* t_3 *(each of (a)−(c) scale adjusted), and (d)* $N(0, 5^2 \times (0.03)^2)$

5. Performance of the proposed method

In order to evaluate the performance of the proposed model, described in Section 2, we conduct a simulation study. This analysis is based on the simulated data sets generated under different settings. For each $m = 100$, 500 and 1000, we generated 100 data sets. Here we set $r = 2$, $x = (1, x_1)^T$ and generate m copies of x_1 from $\mathrm{N}(10, (\sqrt{2})^2)$. For each choice of m, the set of covariates is generated exactly once and used for all 100 data sets. Our choice of β is $\beta = (20, 1)^T$. The sampling error e_i's are generated from $N(0, D_i)$, $i = 1, \ldots, m$, where D_i's are from the set $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$, and each value in the set is allocated to the same number of small areas. Random effects in model (1.1) are generated under three different settings:

$$v_i \sim N(0, 1^2), \tag{5.1}$$

$$v_i \sim (1 - \delta_i)N(0, 1^2) + \delta_i N(0, 5^2), \text{ and} \tag{5.2}$$

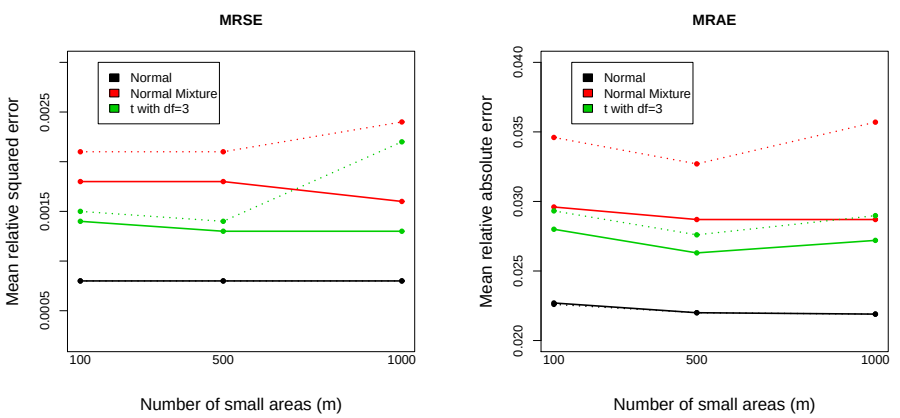
$$v_i \sim t_3, \tag{5.3}$$

where $i = 1, \ldots, m$. For the normal-mixture setup (5.2), we set $\delta_i = 1$ for each i multiple of 5 and keep the rest of the $\delta_i = 0$, the simulated data sets contain 20% of observations from the normal distribution with a variance of 25. Based on the generated set of v_i's , we compute both the θ_i's and y_i's by (1.1). For each of 100 simulated data sets for each setting, we predict θ_i's based on the Fay-Herriot model and the proposed area-level normal-mixture model. We measure the performance of each prediction method by computing the (empirical) mean squared error (MSE)= $\frac{1}{m}\sum_{i=1}^{m}(\theta_i - \hat{\theta}_i)^2$, the mean absolute error (MAE)=$\frac{1}{m}\sum_{i=1}^{m}|\theta_i - \hat{\theta}_i|$, the mean relative squared error (MRSE)= $\frac{1}{m}\sum_{i=1}^{m}\frac{(\theta_i - \hat{\theta}_i)^2}{\theta_i^2}$ and the mean relative absolute error (MRAE)= $\frac{1}{m}\sum_{i=1}^{m}\frac{|\theta_i - \hat{\theta}_i|}{\theta_i}$, where θ_i's are true and $\hat{\theta}_i$'s are estimated small area means (for our simulation setup, all the θ_i's are positive). These empirical deviation measures are typically used in the small area estimation literature to compare the accuracy of various estimation methods (Rao, 2003). For each simulated dataset, we compute MSE, MAE, MRAE and MRSE for two different methods and report the average values based on all simulated data sets. The results of the simulation study are presented in Tables 2 and 3. In Table 2 we report the MSE and MAE and in Figure 4

we plot the MRAE and MRSE based on the overall simulation study. Table 3 shows a more detailed result when the v_i's are drawn according to equation (5.2). From Table 3 we can compare the performance of the two prediction methods for outlying areas (random effects drawn from $N(0,5^2)$) and non-outlying areas (random effects drawn from $N(0,1^2)$), separately. The simulation results indicate that the proposed method tends to perform better than the Fay-Herriot method when the possibility of the presence of outliers is high, and performs similarly otherwise.

Table 2: Comparison of the methods based on the simulated MSE and MAE of prediction. The results are based on 100 simulated data sets

		m=100		m=500		m=1000	
Scenario		Proposed	FH	Proposed	FH	Proposed	FH
(5.1) Normal	MSE	0.72	0.71	0.69	0.69	0.68	0.68
	MAE	0.67	0.67	0.66	0.66	0.66	0.65
(5.2) Mixture	MSE	1.48	1.75	1.49	1.81	1.30	1.87
	MAE	0.86	1.01	0.85	0.98	0.84	1.04
(5.3) t_3	MSE	1.14	1.27	1.01	1.20	1.14	1.30
	MAE	0.83	0.84	0.79	0.81	0.80	0.84



(a) (b)

Figure 4: *(a) The mean relative squared error (MRSE) and (b) the mean relative absolute error (MRAE) based on* 100 *simulated data sets; A dotted line for the Fay-Herriot method and a solid line for the proposed method.*

Table 3: Comparison of the methods based on the simulated MSE, MAE, MRSE and MRAE of prediction. The results are based on 100 simulated data sets. The performance of the methods is compared separately for outlying and non-outlying areas based on the simulation design.

		Scenario (5.2) Mixture					
		m=100		m=500		m=1000	
		Proposed	FH	Proposed	FH	Proposed	FH
MSE	$A_1 = 1^2$	0.90	1.26	0.80	1.06	0.80	1.32
	$A_2 = 5^2$	3.39	3.69	4.25	4.80	3.28	4.03
MAE	$A_1 = 1^2$	0.73	0.88	0.69	0.82	0.70	0.91
	$A_2 = 5^2$	1.43	1.47	1.49	1.61	1.39	1.59
100×MRSE	$A_1 = 1^2$	0.10	0.14	0.09	0.12	0.09	0.15
	$A_2 = 5^2$	0.43	0.50	0.53	0.56	0.44	0.61
10×MRAE	$A_1 = 1^2$	0.25	0.30	0.23	0.27	0.24	0.30
	$A_2 = 5^2$	0.50	0.52	0.51	0.54	0.49	0.57

6. Discussion

In this paper, we propose a robust alternative to the Fay-Herriot model. The proposed hierarchical Bayesian estimation procedure is straightforward. Another robust alternative is a t-distribution for the random effects, which requires information regarding the degrees of freedom. Xie et al. (2007) proposed a method to estimate the degrees of freedom. However, Bell and Huang pointed out that only a very limited information could be extracted from the data regarding the degrees of freedom parameter. We propose a method based on non-informative priors for the parameters. We provide sufficient conditions for the propriety of the resulting posterior distributions.

Model-based small area estimates depend on the accuracy of the underlying model assumptions. Larger values of the area specific random effects may be caused by a poor choice of the linking model or the lack of predictive quality of the auxiliary variables. If the model-based estimates of the area specific random effects are significantly larger for some areas compared to the other areas, it is probably meaningful to retain the direct estimates instead of the model-based estimates for those areas to avoid possible inaccuracy. Nevertheless, we should be cautious in this recommendation if there is any indication that the sampling variance is underestimated.

Datta and Lahiri (1995) recommended heavy-tailed priors for random effects by emphasizing the fact that estimators obtained by using these priors were similar to direct estimators for the areas with extreme observations. However, the estimators for non-outlying areas should shrink direct estimators more towards synthetic estimators. Also, the magnitude of this shrinkage may depend on the quality of the auxiliary information. While for an outlying observation our model limits the shrinkage of the Bayes predictor to the synthetic estimator, for non-outlying observations it enables the Bayes predictors to retain the shrinkage to the synthetic estimator when the regression model provides a good fit.

Acknowledgments

The authors would like to thank to Dr. Jerry Maples from the Census Bureau for providing and explaining the poverty data in our application. The authors are also grateful to him and to Dr. William R. Bell for an internal review of an earlier version of the manuscript. Many of their valuable comments, particularly some substantive comments by Dr. Bell, led to a significantly improved manuscript. The research by A. Mandal is partially supported by the NSA Grant H98230-13-1-0251.

REFERENCES

1. ANGERS, J. F., BERGER, J. O., (1991). Robust hierarchical Bayes estimation of exchangeable means. *Canadian Journal of Statistics*, **19**, 39–56.

2. BELL, W. R., HUANG, E. T., (2006). Using t-distribution to deal with outliers in small area estimation. *Proceedings of Statistics Canada Synposium 2006 Methodological issues in measuring population health*.

3. BERGER, J. O., (1984). The robust Bayesian viewpoint (with discussion). In *Robustness of Bayesian Analysis*, Ed.: J. Kadane, North-Holland, Amsterdam.

4. CHAMBERS, R., CHANDRA, H., SALVATI, N., TZAVIDIS, N., (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society, Series B:*, **76**, 47–69.

5. DATTA, G. S., (2009). Model-based approach to small area estimation. *Handbook of Statistics: Sample Surveys: Inference and Analysis*, *29B*, Eds.: D. Pfeffermann and C. R. Rao, The Netherlands: North-Holland. 251–288.

6. DATTA, G. S., GHOSH, M., (2012). Small area shrinkage estimation. *Statistical Science*, **27**, 95–114.

7. DATTA, G. S., HALL, P. G., MANDAL, A., (2011). Model selection by testing for the presence of small-area effects in area-level data. *Journal of the American Statistical Association*, **106**, 362–374.

8. DATTA, G. S., LAHIRI, P., (1995). Robust hierarchical Bayesian estimation of small area characteristics in presence of covariates and outliers. *Journal of Multivariate Analysis*, **54**, 310–328.

9. DATTA, G. S., MANDAL, A., (2015). Small area estimation with uncertain random effects. To appear in *Journal of the American Statistical Association*, **110**, DOI: 10.1080/01621459.2015.1016526.

10. DATTA, G. S., RAO, J. N. K., SMITH, D., (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, **92,** 183–196.

11. DEY, D. K., BERGER, J. O., (1983). On truncation of shrinkage estimators in simultaneous estimation of normal means. *Journal of the American Statistical Association*, **78**, 865–869.

12. EFRON, B., MORRIS, C., (1971). Limiting the risk of Bayes and empirical Bayes estimators, Part I: The Bayes Case. *Journal of the American Statistical Association*, **67**, 130–139.

13. FAY, R. E., 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**, 269-277.

14. GHOSH, M., RAO, J. N. K., (1994). Small area estimation: an appraisal. *Statistical Science*, **9**, 55–93.

15. GHOSH, M., MAITI, T., ROY, A., (2008). Influence functions and robust Bayes and empirical Bayes small area estimation. *Biometrika*, **95**, 573–585.

16. JIANG, J., LAHIRI, P., (2006). Mixed model prediction and small area estimation. *Test*, **15**, 1–96.

17. PFEFFERMANN, D., (2013). New important developments in small area estimation. *Statistical Science*, **28**, 40–68.

18. PRASAD, N. G. N., RAO, J. N. K., (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163–171.

19. RAO, J. N. K., (2003). *Small Area Estimation*. Wiley-Interscience, Hoboken, NJ.

20. RAO, J. N. K., (2011). Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal. *Statistical Science*, **26**, 240–256.

21. SCOTT, J. G., BERGER, J. O., (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, **136**, 2144–2162.

22. SINHA, S. K., RAO, J. N. K., (2009). Robust small area estimation. *The Canadian Journal of Statistics*, **37**, 381–399.

23. STEIN, C. M., (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1954–1955, **I**, 197-206. Univ. California Press, Berkeley.

24. STEIN, C. M., (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**, 1135–1151.

25. XIE, D., RAGHUNATHAN, T. E., LEPKOWSKI, J. M., (2007). Estimation of the proportion of overweight individuals in small areas - a robust extension of the Fay-Herriot model. *Statistics in Medicine*, **26**, 2699–2715.

Appendix

Gibbs sampling for the proposed model

In order to apply our model, we use Gibbs sampling. We derive the set of full conditional distributions from the posterior joint density of $\theta = (\theta_1, \ldots, \theta_m)^T$, $\beta = (\beta_1, \ldots, \beta_r)^T$, $\delta = (\delta_1, \ldots, \delta_m)^T$, A_1, A_2 and p, which is given by

$$\begin{aligned}\pi(\theta,\beta,A_1,A_2,\delta,p|y) \propto & \left\{\prod_{i=1}^{m}\exp\left\{-\frac{(y_i-\theta_i)^2}{2D_i}\right\}\right\}\prod_{i=1}^{m}\left[p^{\delta_i}(1-p)^{1-\delta_i}\right. \\ & \left\{\frac{1}{\sqrt{A_1}}\times\exp\left\{-\frac{(\theta_i-x_i^T\beta)^2}{2A_1}\right\}\right\}^{\delta_i} \\ & \left.\times\left\{\frac{1}{\sqrt{A_2}}\times\exp\left\{-\frac{(\theta_i-x_i^T\beta)^2}{2A_2}\right\}\right\}^{1-\delta_i}\right] \\ & \times A_1^{-\alpha_1}A_2^{-\alpha_2}\times I(0<A_1<A_2). \end{aligned} \tag{6.1}$$

From (6.1), we get the following full conditional distributions:

(I) $\theta_i|\beta,A_1,A_2,\delta,p,y \overset{\text{ind}}{\sim} N\Big(\frac{D_i x_i^T\beta + A_{2-\delta_i}y_i}{D_i + A_{2-\delta_i}}, \frac{D_i A_{2-\delta_i}}{D_i + A_{2-\delta_i}}\Big)$, $i = 1, \ldots, m$;

(II) $\beta|\theta,A_1,A_2,\delta,p,y \sim N\left(G^{-1}\left[\sum_{i=1}^{m}A_{2-\delta_i}^{-1}x_i\theta_i\right], G^{-1}\right)$, where G is given by $\sum_{i=1}^{m}A_{2-\delta_i}^{-1}x_ix_i^T$;

(III) $p|\theta,\beta,A_1,A_2,\delta,y \sim Beta\left(\sum_{i=1}^{m}\delta_i+1, m-\sum_{i=1}^{m}\delta_i+1\right)$;

(IV) $A_1|A_2,\theta,\beta,\delta,p,y$ has the pdf $f_1(A_1)$, where,

$$f_1(A_1) \propto A_1^{-(\alpha_1+\sum_{i=1}^{m}\frac{\delta_i}{2})}\exp\left\{-\sum_{i=1}^{m}\frac{\delta_i(\theta_i-x_i^T\beta)^2}{2A_1}\right\}I(A_1<A_2),$$

(V) $A_2|A_1,\theta,\beta,\delta,p,y$ has the pdf $f_2(A_2)$, where,

$$f_2(A_2) \propto A_2^{-(\alpha_2+\sum_{i=1}^{m}\frac{(1-\delta_i)}{2})}\exp\left\{-\sum_{i=1}^{m}\frac{(1-\delta_i)(\theta_i-x_i^T\beta)^2}{2A_2}\right\}I(A_1<A_2),$$

(VI) For $i = 1, \ldots, m$, $\delta_i|\theta,\beta,A_1,A_2,p,y$ are independent with

$$P(\delta_i=1|\theta,\beta,p,y) = \frac{\frac{p}{\sqrt{A_1}}\exp\left\{-\frac{(\theta_i-x_i^T\beta)^2}{2A_1}\right\}}{\frac{p}{\sqrt{A_1}}\exp\left\{-\frac{(\theta_i-x_i^T\beta)^2}{2A_1}\right\}+\frac{(1-p)}{\sqrt{A_2}}\exp\left\{-\frac{(\theta_i-x_i^T\beta)^2}{2A_2}\right\}}.$$

Our goal is to estimate θ_i, i.e., small area mean for the i^{th} area, $i = 1,\ldots,m$. We implement Gibbs sampling using the conditional distributions (I)−(VI) in order to find posterior means and standard deviations of θ_i's. Conditional distribution (IV) and (V) may not have always admit a closed form expression.

Proof of Theorem 2.1

Note that under the mixture model, the likelihood function of the model parameters β, A_1, A_2 and p based on the marginal distribution of $y_1,\ldots,y_m$ is given by

$$L(\beta,A_1,A_2,p) = C \times \prod_{i=1}^{m} \left[\frac{p}{(A_1+D_i)^{\frac{1}{2}}} e^{-\frac{(y_i - x_i^T\beta)^2}{2(A_1+D_i)}} + \frac{(1-p)}{(A_2+D_i)^{\frac{1}{2}}} e^{-\frac{(y_i - x_i^T\beta)^2}{2(A_2+D_i)}} \right], \tag{6.2}$$

where C is a generic positive constant not depending on the model parameters. Suppose for $0 < a < b < \infty$ we have $a \le D_i \le b$, $i = 1,\ldots,m$. Since $(A_1+b) \ge (A_1+D_i) \ge (a/b)(A_1+b)$, $(A_2+b) \ge (A_2+D_i) \ge (a/b)(A_2+b)$, from (6.2)

$$L(\beta,A_1,A_2,p) \le C \times \prod_{i=1}^{m} \left[\frac{p}{(A_1+b)^{\frac{1}{2}}} e^{-\frac{(y_i - x_i^T\beta)^2}{2(A_1+b)}} + \frac{(1-p)}{(A_2+b)^{\frac{1}{2}}} e^{-\frac{(y_i - x_i^T\beta)^2}{2(A_2+b)}} \right]. \tag{6.3}$$

For $k = 0,1,\cdots,m$, let $P_k = \{S_1^{(k)}, S_2^{(k)}\}$ be an arbitrary partition of $\{1,2,\cdots,m\}$, where $S_1^{(k)}$ has k elements and $S_2^{(k)}$ has $m-k = l(\text{say})$ elements. Let $\mathscr{P}_k$ denote all $\binom{m}{k}$ collections of $\{S_1^{(k)}, S_2^{(k)}\}$. Then, expanding the product of the right hand side of (6.3), we get

$$L(\beta,A_1,A_2,p) \le C \sum_{k=0}^{m} \sum_{P_k \in \mathscr{P}_k} \frac{p^k (1-p)^{m-k} e^{-\sum_{i\in S_1^{(k)}} \frac{(y_i - x_i^T\beta)^2}{2(A_1+b)} - \sum_{i\in S_2^{(k)}} \frac{(y_i - x_i^T\beta)^2}{2(A_2+b)}}}{(A_1+b)^{\frac{k}{2}} (A_2+b)^{\frac{m-k}{2}}}. \tag{6.4}$$

To show propriety of the posterior density, we show integrability of each of the 2^m summands on the right hand side of (6.4) with respect to the prior given in (2.2).

We first consider the case $k = 0$. Here $\mathscr{P}_0$ has one element and $S^{(0)}$ is a null set. Let

$Q(y) = y^T[I - X(X^TX)^{-1}X^T]y$. In this case, the integral $I^{(0)}$ of the term is

$$\begin{aligned} I^{(0)} &= C\int_0^\infty \int_{R^r} \int_0^{A_2} \int_0^1 (1-p)^m dp \frac{dA_1}{A_1^{\alpha_1}} \frac{A_2^{-\alpha_2}}{(A_2+b)^{-\frac{m}{2}}} e^{-\Sigma_{i=1}^m \frac{(y_i - x_i^T\beta)^2}{2(A_2+b)}} d\beta dA_2 \\ &= C\int_0^\infty A_2^{1-\alpha_1-\alpha_2}(A_2+b)^{-\frac{m}{2}} e^{-\frac{1}{2}\frac{Q(y)}{A_2+b}} dA_2 \qquad (\text{since } \alpha_1 < 1) \\ &\leq C\int_0^\infty A_2^{1-\alpha_1-\alpha_2}(A_2+b)^{-\frac{m-r}{2}} dA_2 < \infty, \end{aligned} \tag{6.5}$$

if and only if $2-\alpha_1-\alpha_2 > 0$ and $1-\alpha_1-\alpha_2-\frac{m-r}{2} < -1$, which are equivalent to the conditions outlined in Theorem 2.1.

For the case $k = m$, again there is one term in $\mathscr{P}_m$ and the resulting integral, proceeding as in $I^{(0)}$, is bounded above by

$$\begin{aligned} & C\int_0^\infty A_1^{-\alpha_1}(A_1+b)^{-\frac{m-r}{2}} \int_{A_1}^\infty A_2^{-\alpha_2} dA_2 dA_1 \\ = \; & C\int_0^\infty A_1^{1-\alpha_1-\alpha_2}(A_1+b)^{-\frac{m-r}{2}} dA_1 \;\; (\text{since } \alpha_2 > 1) < \infty, \end{aligned} \tag{6.6}$$

under the conditions of the theorem.

Now consider a case where $1 \leq k \leq m-1$. Let $S_1^{(k)}$ be a set of indices $\{i_1, \ldots, i_k\}$ and let $S_2^{(k)} = \{j_1, \ldots, j_l\} = \{1, 2, \cdots, m\} \setminus S_1^{(k)}$. Let us define, $M_1 = (x_{i_1}, \ldots, x_{i_k})^T$ and $M_2 = (x_{j_1}, \ldots, x_{j_l})^T$. Suppose $g = \text{rank}(M_1)$. If $g > 0$, suppose $B \equiv \{\alpha_1, \ldots, \alpha_g\} \subset \{i_1, \ldots, i_k\}$, so that $\{x_{\alpha_1}, \ldots, x_{\alpha_g}\}$ is linearly independent. If $g = 0$, the set B is empty. Suppose $\{\gamma_1, \ldots, \gamma_{r-g}\} \subset \{j_1, \ldots, j_l\}$ such that $\{x_{\alpha_1}, \ldots, x_{\alpha_g}, x_{\gamma_1}, \ldots, x_{\gamma_{r-g}}\}$ is linearly independent. Let us define the $r \times r$ matrix $F = \left(x_{\alpha_1}, \ldots, x_{\alpha_g}, x_{\gamma_1}, \ldots, x_{\gamma_{r-g}}\right)^T$, which is non-singular. Consider the non-singular linear transformation of β by $\phi = F\beta$. With these developments, the integral of the term identified by $\{S_1^{(k)}, S_2^{(k)}\}$ in the right hand side of (6.4) with respect to the prior $\pi(\beta, A_1, A_2, p)$ is bounded above by a positive generic constant C times

$$\int_0^\infty \int_{A_1}^\infty \int_{R^r} \frac{A_1^{-\alpha_1} A_2^{-\alpha_2} e^{-\Sigma_{i\in S_1^{(k)}} \frac{(y_i - x_i^T\beta)^2}{2(A_1+b)} - \Sigma_{i\in S_2^{(k)}} \frac{(y_i - x_i^T\beta)^2}{2(A_2+b)}}}{(A_1+b)^{\frac{k}{2}}(A_2+b)^{\frac{l}{2}}} d\beta dA_2 dA_1$$

$$
\begin{aligned}
&\leq \int_0^\infty \int_{A_1}^\infty \int_{R^r} \frac{A_1^{-\alpha_1} A_2^{-\alpha_2} e^{-\sum_{u=1}^{g} \frac{(y_{\alpha_u} - x_{\alpha_u}^T \beta)^2}{2(A_1+b)} - \sum_{t=1}^{r-g} \frac{(y_{\gamma_t} - x_{\gamma_t}^T \beta)^2}{2(A_2+b)}}}{(A_1+b)^{\frac{k}{2}} (A_2+b)^{\frac{l}{2}}} d\beta dA_2 dA_1 \\
&= \int_0^\infty \int_{A_1}^\infty \int_{R^r} \frac{A_1^{-\alpha_1} A_2^{-\alpha_2} e^{-\sum_{u=1}^{g} \frac{(y_{\alpha_u} - \phi_u)^2}{2(A_1+b)} - \sum_{t=1}^{r-g} \frac{(y_{\gamma_t} - \phi_{g+t})^2}{2(A_2+b)}}}{(A_1+b)^{\frac{k}{2}} (A_2+b)^{\frac{l}{2}}} d\phi dA_2 dA_1 \\
&= \int_0^\infty \int_{A_1}^\infty \frac{A_1^{-\alpha_1} A_2^{-\alpha_2}}{(A_1+b)^{\frac{k-g}{2}} (A_2+b)^{\frac{l-r+g}{2}}} dA_1 dA_2 \\
&\leq \int_0^\infty \int_{A_1}^\infty \frac{A_1^{-\alpha_1} A_2^{-\alpha_2}}{(A_1+b)^{\frac{k-g}{2}} (A_1+b)^{\frac{l-r+g}{2}}} dA_2 dA_1 \\
&= \int_0^\infty \frac{A_1^{1-\alpha_1-\alpha_2}}{(A_1+b)^{\frac{k-g}{2}} (A_1+b)^{\frac{l-r+g}{2}}} dA_1 \\
&= \int_0^\infty \frac{A_1^{1-\alpha_1-\alpha_2}}{(A_1+b)^{\frac{m-r}{2}}} dA_1 < \infty, \qquad (6.7)
\end{aligned}
$$

by the conditions of the theorem. Since the integrability conditions do not depend k or on the indices $\{i_1, \ldots, i_k\}$ and $\{j_1, \ldots, j_l\}$ and on the values k and l, the conditions $2-\alpha_1-\alpha_2 > 0$ and $m > r+2(2-\alpha_1-\alpha_2)$ will be sufficient to ensure the propriety of the posterior. □

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 91–104

VARIATIONAL APPROXIMATIONS FOR SELECTING HIERARCHICAL MODELS OF CIRCULAR DATA IN A SMALL AREA ESTIMATION APPLICATION

Daniel Hernandez-Stumpfhauser[1], **F. Jay Breidt**[2], **Jean D. Opsomer**[3]

ABSTRACT

We consider hierarchical regression models for circular data using the projected normal distribution, applied in the development of weights for the Access Point Angler Intercept Survey, a recreational angling survey conducted by the US National Marine Fisheries Service. Weighted estimates of recreational fish catch are used in stock assessments and fisheries regulation. The construction of the survey weights requires the distribution of daily departure times of anglers from fishing sites, within spatio-temporal domains subdivided by the mode of fishing. Because many of these domains have small sample sizes, small area estimation methods are developed. Bayesian inference for the circular distributions on the 24-hour clock is conducted, based on a large set of observed daily departure times from another National Marine Fisheries Service study, the Coastal Household Telephone Survey. A novel variational/Laplace approximation to the posterior distribution allows fast comparison of a large number of models in this context, by dramatically speeding up computations relative to the fast Markov Chain Monte Carlo method while giving virtually identical results.

Key words: deviance information criterion, Laplace approximation, model selection, projected normal distribution.

1. Introduction

In the United States, the Marine Recreational Fisheries Statistics Survey (MRFSS) has been the traditional source of information on recreational fishing in saltwater. The key question for stock assessment and fisheries regulation is the amount of recreational fishing catch, determined from the simple relationship

$$\text{(recreational catch)} = \text{(catch per angler-trip)} \times \text{(number of angler-trips)}.$$

[1]University of North Carolina–Chapel Hill. E-mail: danielhs@live.unc.edu
[2]Colorado State University. E-mail: jbreidt@stat.colostate.edu
[3]Colorado State University. E-mail: jopsomer@stat.colostate.edu

Due to a number of coverage and measurement issues, the two factors in the above expression are measured using different surveys: (catch per angler-trip) is measured by an on-site survey called the Access Point Angler Intercept Survey (APAIS), while the number of angler-trips is measured by an off-site survey called the Coastal Household Telephone Survey (CHTS). Data from these two surveys are combined to estimate the recreational catch in 17 US states along the coast of the Atlantic Ocean and the Gulf of Mexico, during six two-month waves (January–February, March–April,..., November–December), in four different fishing modes (from the shoreline, from a private boat, from a small guided vessel called a charter boat, or from a large guided vessel called a party boat). Because the state of Florida is divided into its Atlantic coast and its Gulf of Mexico coast, we will refer to 18 "states" instead of 17.

As part of the weighting procedure for the APAIS, estimates are needed for the fraction of anglers who leave the fishing site during a prespecified time interval on a selected day. In principle, these estimates could be readily obtained from extensive historical data from the CHTS, consisting of reports on 980,000 trips between 1990 and 2008. These data include the angler's departure time (on a 24-hour clock) from the fishing site, the mode of fishing, the fishing date (from which we determine the two-month wave), and the fishing site (from which we determine the state). Figure 1 shows these data in histogram form for the state of Alabama. There are 24 histograms, corresponding to six waves by four fishing modes. The bars in the histograms, when normalized by sample sizes, can be regarded as direct estimates $\widehat{F}^{\text{direct}}_{hijk}$ of the hourly fractions of daily departures by state, wave, and mode:

$$F_{hijk} = \text{fraction of a day's anglers leaving a site during hour } h \text{ in state } i, \text{ wave } j, \text{ mode } k.$$

The fraction for any prespecified block of hours is then modeled as $\sum_h F_{hijk}$, where the sum is over all hours h in that block. Other time intervals are rounded to the nearest whole hours, for simplicity.

The direct estimates $\widehat{F}^{\text{direct}}_{hijk}$ from the off-site CHTS data are unbiased, but have a small (or even zero) sample size in many of the (h,i,j,k) cells, of which there are

$$(24 \text{ hours}) \times (18 \text{ states}) \times (6 \text{ waves}) \times (4 \text{ modes}) = (10368 \text{ cells}).$$

We therefore consider the small area estimation approach, combining the direct estimates with modeled estimates using the Fay and Herriot (1979) estimation method-

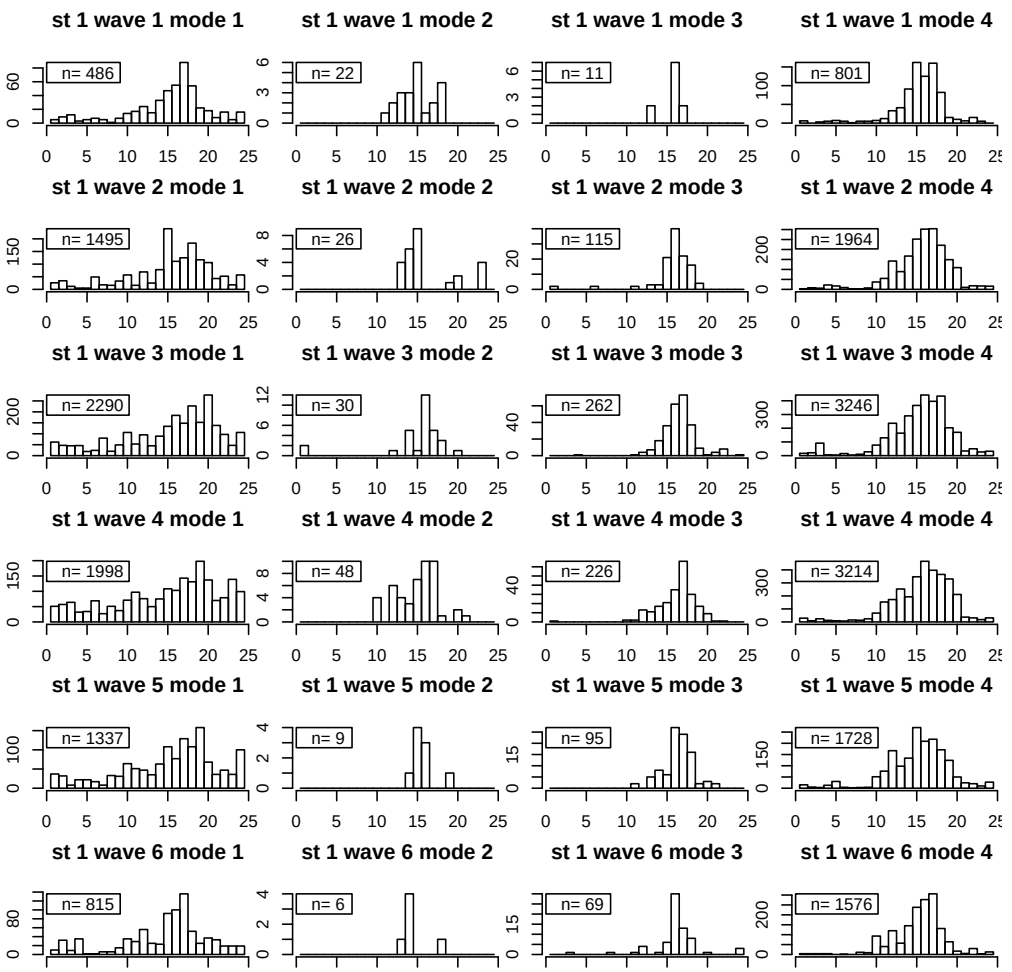


Figure 1: Histograms of trip departure times from the Coastal Household Telephone Survey for the state of Alabama (st 1) in six waves (top row = wave 1 = January–February, ..., bottom row = wave 6 = November–December) and four modes (column 1 = shoreline, 2 = private boat, 3 = charter boat, 4 = party boat).

ology. Briefly, we consider an area-level linear mixed model

$$\widehat{F}_{hijk}^{\text{direct}} = F_{hijk} + e_{hijk} = F_{hijk}^{\text{model}} + u_{hijk} + e_{hijk}$$

for $h = 1, \ldots, 23$ hours, where the sampling errors are assumed to be

$$\mathbf{e}_{ijk} = (e_{1ijk}, e_{2ijk}, \ldots, e_{23,ijk})^T \sim \text{independent } \mathcal{N}(0, \Psi_{ijk}),$$

with Ψ_{ijk} known, and where the model errors are assumed to be

$$\mathbf{u}_{ijk} = (u_{1ijk}, u_{2ijk}, \ldots, u_{23,ijk})^T \sim \text{independent } \mathcal{N}(0, \sigma^2\Delta_{ijk}),$$

with Δ_{ijk} of known form. Sampling errors and model errors are assumed to be independent. To implement the estimation strategy, we replace Ψ_{ijk} by design-based variance estimates and we choose Δ_{ijk} to be the variance of a scaled multinomial random vector, specified as follows. Consider a vector of 24 independent normal random variables with covariance matrix

$$\begin{aligned}\sigma^2 \text{diag}\left\{G_{1ijk}^{\text{model}}, \ldots, G_{23ijk}^{\text{model}}, G_{24ijk}^{\text{model}}\right\} \\ = \quad \sigma^2 \text{diag}\left\{F_{1ijk}^{\text{model}}\left(1 - F_{1ijk}^{\text{model}}\right), \ldots, F_{24ijk}^{\text{model}}\left(1 - F_{24ijk}^{\text{model}}\right)\right\}.\end{aligned}$$

Then $\sigma^2\Delta_{ijk}$ is the covariance matrix of the first 23 elements of the vector, conditioned on the sum of the 24 elements being equal to one; namely,

$$\begin{aligned}\sigma^2\Delta_{ijk} \quad = \quad & \sigma^2 \text{diag}\left\{G_{1ijk}^{\text{model}}, \ldots, G_{23ijk}^{\text{model}}\right\} \\ & - \frac{\sigma^2}{\sum_{\tau=1}^{24} G_{\tau ijk}^{\text{model}}} \begin{bmatrix} G_{1ijk}^{\text{model}} \\ \vdots \\ G_{23ijk}^{\text{model}} \end{bmatrix} \left[G_{1ijk}^{\text{model}}, \cdots, G_{23ijk}^{\text{model}}\right]. \qquad (1)\end{aligned}$$

We use a projected normal model for F_{hijk}^{model} to account for the circular nature of the time-of-day departure data, replacing F_{hijk}^{model} by posterior means $\text{E}\left[F_{hijk}^{\text{model}} \mid D\right]$ and also G_{hijk}^{model} by $\left(\text{E}\left[F_{hijk}^{\text{model}} \mid D\right]\right)\left(1 - \text{E}\left[F_{hijk}^{\text{model}} \mid D\right]\right)$ for implementation. The mean vector in the projected normal includes state, wave, and mode effects to account for the spatial and temporal distribution of fishing behavior. Since we consider various interactions among the effects as well as placement within the hierarchy (essentially, specifying whether a given effect is treated as fixed or random), we are interested in conducting model selection.

The main contribution of the present paper is to show that in this small area

estimation context, with a model somewhat more complex than a hierarchical linear model (due to the embedding in a projected normal model), fast and accurate model selection can be accomplished with a Laplace/variational approximation. Specifically, we show that a simple and fast deterministic approximation can replace a sophisticated Markov Chain Monte Carlo (MCMC) sampler, giving results that are essentially identical at a far lower computational cost. In this paper, we emphasize model selection as both the motivation for the deterministic approximation and the evaluation of its accuracy. However, the Laplace/variational approximation can also be used effectively in model estimation and inference even when no model selection is needed.

In §2.1, we briefly review the projected normal distribution. The MCMC procedure that serves as the benchmark for comparison is presented in §2.2. The variational approximation is given in §3.1 with its Laplace refinement in §3.2. Model selection criteria based on MCMC and on the Laplace/variational approximation are compared in §4; discussion follows in §5.

2. Inference for the projected normal distribution

2.1. The projected normal distribution

Suppose $X = (X_1, X_2)^T \sim \mathcal{N}(\mu, I_2)$, the bivariate normal distribution with mean vector μ and identity covariance matrix I_2. Writing X in polar coordinates, we have

$$X_1 = \|X\| \cos D = R \cos D, \quad X_2 = \|X\| \sin D = R \sin D.$$

Discarding the random length $R \in (0, \infty)$, the random angle $D \in [0, 2\pi)$ has a projected normal distribution, $\mathscr{PN}(\mu, I_2)$. As illustrated in Figure 2, the parameter vector μ plays the role of both "location" and "spread" for the projected normal: the further μ lies from the origin, the more concentrated the $\mathscr{PN}$ distribution around the direction determined by μ. As $\mu \to 0$, the $\mathscr{PN}$ distribution converges to the uniform distribution on the unit circle. In our application, the departure time d_{ijkt} for trip t in state i, wave j, mode k is on the 24-hour clock. Converting clock time to $[0, 2\pi)$, we model $D_{ijkt} = 2\pi d_{ijkt}/24$ as independent and identically distributed projected normals within state×wave×mode cells. For observations following a projected normal distribution, the fraction F_{hijk} for a given hour h is the integral of the projected normal probability density function over the interval $(2\pi(h-1)/24, 2\pi h/24]$.

Presnell, Morrison and Littell (1998) used the projected normal distribution as the basis for the Spherically Projected Multivariate Linear Model (SPMLM) for

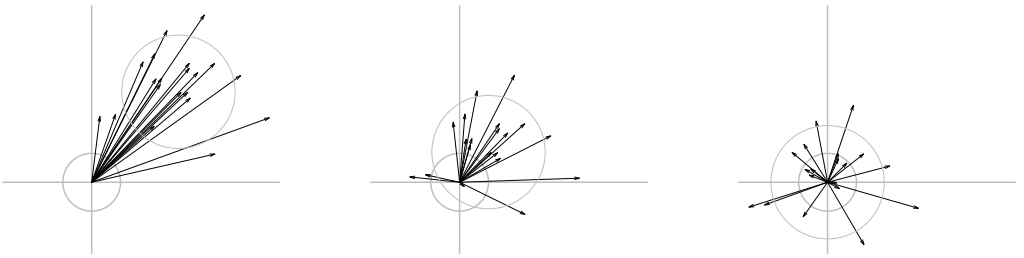


Figure 2: Realizations ($n = 20$) from three projected normal distributions. The large circle is centered at mean vector μ of bivariate normal $\mathcal{N}(\mu, I_2)$ and contains 95% of its probability. Arrows are the realized bivariate normal random vectors $(R\cos D, R\sin D)$. Projected normal random variables are the angles D, or the intersections of the normal random vectors with the unit circle (small circle), scaled to $[0, 2\pi)$. Left: Projected normal distribution with mode equal to $\pi/4$ and with low variance. Middle: Projected normal distribution with mode equal to $\pi/4$ and with high variance. Right: Projected normal distribution that is uniform on the unit circle.

directional data, specifying μ as a linear model. Parameters of the model were estimated with the maximum likelihood and the EM algorithm in Presnell et al. (1998). In the current paper, we specify hierarchical linear models for μ_{ijk} in terms of categorical covariates for the state, wave and mode. We conduct Bayesian inference for the model, comparing approximate posterior inference based on Markov Chain Monte Carlo to approximate inference based on deterministic approximations.

2.2. Markov Chain Monte Carlo for the projected normal distribution

The key step in conducting Bayesian inference under the SPMLM is to augment the observed angles $\{D_{ijkt}\}$ with the latent lengths $\{R_{ijkt}\}$, so that the structure of the complete data is simply that of a normal linear model. See Nuñez-Antonio and Gutiérrez-Peña (2005), Nuñez-Antonio, Gutiérrez-Peña, and Escalera (2011), and Hernandez-Stumpfhauser (2012) for details.

The likelihood for the complete-data model is the product of the joint densities of $\left(R_{ijkt}, D_{ijkt}\right)$ which can be obtained by a change of variables $X_{ijkt} = R_{ijkt}A_{ijkt}$, where X_{ijkt} is distributed as $\mathcal{N}\left(\mu_{ijk}, I_2\right)$ and $A_{ijkt} = \left(\cos\left(D_{ijkt}\right), \sin\left(D_{ijkt}\right)\right)^T$:

$$p\left(R_{ijkt}, D_{ijkt} \mid \mu_{ijk}\right) = \frac{1}{2\pi} r_{ijkt} \exp\left\{-\frac{1}{2}\left(R_{ijkt}A_{ijkt} - \mu_{ijk}\right)^T \left(R_{ijkt}A_{ijkt} - \mu_{ijk}\right)\right\}.$$

We specify conjugate normal priors for μ_{ijk}. For example, for a model specified as $\mu_{ijk} = \mu + m_k + s_i + w_j$, we set vague normal priors for the overall mean μ and mode effects m_k, and mean-zero normal priors with inverse gamma variances for the random state effects s_i and wave effects w_j.

In this work, we draw the latent lengths using a slice sampler (Neal 2003). Given the latent lengths and the conjugate priors, the full conditionals of the model parameters all have closed forms, and so the Gibbs sampler is fast and easy to conduct. Nonetheless, the large number of models to be evaluated led us to consider fast, deterministic approximations to the posterior distribution. This is the subject of the next section.

3. Deterministic approximations to the posterior

3.1. Variational approximation

In this context, a carefully-developed MCMC works well and serves as a benchmark for comparison. But it is extremely slow, given the very large size of the off-site CHTS data set. Because we wanted to compare a number of different model specifications, we investigated replacing the MCMC approximation of the full posterior distribution by a deterministic "variational approximation" that is easier to compute.

The variational idea is to find the best approximation of the posterior within a class of densities $\mathscr{Q}$, which is chosen so that the densities in the class are more analytically tractable than the posterior density itself. A natural choice for the "best" approximating density in $\mathscr{Q}$, and the one most commonly used, is the density that minimizes the Kullback-Leibler (KL) distance between it and the posterior density. Let D denote the observed data and ω denote the unknown parameters, so that $p(D,\omega)$ is their joint density and $p(\omega \mid D)$ is the unknown posterior density. Let $q(\omega)$ denote a density in $\mathscr{Q}$. Finding q that minimizes the KL distance to $p(\omega \mid D)$ is equivalent to maximizing the *variational lower bound*, denoted by

$$\underline{p}(D;q) = \exp\left[\int q(\omega)\log\left\{\frac{p(D,\omega)}{q(\omega)}\right\} d\omega\right]. \tag{2}$$

Let

$$q_* = \max_{q\in\mathscr{Q}} \underline{p}(D;q).$$

If $\mathscr{Q} = \{\text{all densities } q\}$, then $q_*(\omega) = p(\omega \mid D)$, the true posterior of ω given D. If $\mathscr{Q}$ is a sufficiently rich class of densities, then q_* should be a good approximation to the true posterior. In practice, the approximation method is necessarily of limited

accuracy, so q_* will not converge to the true posterior if the true posterior $\notin \mathcal{Q}$.

If $\mathcal{Q} = \left\{q : q(\omega) = \prod_{m=1}^{M} q_{\omega_m}(\omega_m)\right\}$, then $q_*(\omega) = \prod_{m=1}^{M} q_{*\omega_m}(\omega_m)$, which is called a "mean field variational approximation" (Bishop 2006; Ormerod and Wand 2010). This approximation can be computed efficiently, even for very large samples. See Ormerod and Wand (2010) for an excellent review. The solution satisfies

$$\begin{aligned} q_{*\omega_1}(\omega_1) &\propto \exp\{\mathrm{E}_{-\omega_1} \log p(\omega_1 \mid \omega_2, \ldots, \omega_M, D)\} \\ q_{*\omega_2}(\omega_2) &\propto \exp\{\mathrm{E}_{-\omega_2} \log p(\omega_2 \mid \omega_1, \omega_3, \ldots, \omega_M, D)\} \\ &\vdots \\ q_{*\omega_M}(\omega_M) &\propto \exp\{\mathrm{E}_{-\omega_M} \log p(\omega_M \mid \omega_1, \ldots, \omega_{M-1}, D)\}, \end{aligned}$$

where $\mathrm{E}_{-\omega_m}[\cdot]$ denotes expectation with respect to all of the variational component distributions except $q_{*\omega_m}$.

In our setting, $q_{*\omega_m}(\cdot)$ and $\mathrm{E}_{-\omega_m}[\cdot]$ have simple parametric forms; iteratively updating the parameters leads to the solution. Convergence is assessed by monitoring the change in the lower bound $\underline{p}(D;q)$ from (2).

For simplicity, we begin by considering $\{D_t\}$ independent and identically distributed $\mathscr{PN}(\mu, I_2)$, with prior $p(\mu) = \mathcal{N}_2\left(\mu_0, \sigma_0^2 I_2\right)$. Denoting the observed data as $A_t^T = (\cos D_t, \sin D_t)$, the mean field variational approximation satisfies

$$q_{*\mu}(\mu) = \mathcal{N}\left(\frac{\mu_0/\sigma_0^2 + \sum_{t=1}^{n} \mathrm{E}_{-r_t}(r_t) A_t}{n + \left(1/\sigma_0^2\right)}, \frac{1}{n + \left(1/\sigma_0^2\right)} I_2\right) \tag{3}$$

$$q_{*r_t}(r_t) \propto r_t \exp\left(-\frac{1}{2} r_t^2 + r_t A_t^T \mathrm{E}_{-\mu}(\mu)\right), \tag{4}$$

where the expectations in these expressions are computed iteratively:

$$\begin{aligned} \mathrm{E}_{-\mu}(\mu) &\leftarrow \frac{\mu_0/\sigma_0^2 + \sum_{t=1}^{n} \mathrm{E}_{-r_t}(r_t) A_t}{n + \left(1/\sigma_0^2\right)} \\ b_t &\leftarrow A_t^T \mathrm{E}_{-\mu}(\mu) \\ \mathrm{E}_{-r_t}(r_t) &\leftarrow b_t + \frac{\sqrt{2\pi}\, \exp(b_t^2/2)\, \Phi(b_t)}{1 + \sqrt{2\pi}\, b_t\, \exp(b_t^2/2)\, \Phi(b_t)}. \end{aligned}$$

The mean field variational approximation yields a highly tractable approximate posterior, and the iterative solution is simple to compute and fast to converge. In fact, it is proved in Hernandez-Stumpfhauser (2012) that the parameter iterations for μ converge to the posterior mode of the parameters of the projected normal distribution, denoted here by $\mu^\dagger$. Extension of the mean field variational approx-

imation to the case of a linear model for μ is straightforward, and involves updates of means and variances of each one of the fixed and random effects as well as updates of the means of inverse variances of the random effects. See Hernandez-Stumpfhauser (2012) for details.

3.2. Refinement via Laplace approximation

While the mean field variational approximation of the previous section is simple and fast, it is not very accurate. Indeed, the approximate posterior variance for μ in (3) depends on the sample size but not on the data, and so cannot be accurate except in simple cases. Our approach to improving the accuracy of the variational approximation is to replace $q_{*\mu}(\mu)$ by a Laplace approximation $\mathcal{N}(\mu^\dagger, V^\dagger)$, where $\mu^\dagger$ is the posterior mode and the covariance matrix $V^\dagger$ is the inverse of minus the Hessian of the log posterior distribution evaluated at the mode,

$$V^\dagger = \left(- \left[\begin{array}{cc} \frac{\partial^2}{\partial \mu_1^2} \log p(\mu \mid D) & \frac{\partial^2}{\partial \mu_1 \mu_2} \log p(\mu \mid D) \\ \frac{\partial^2}{\partial \mu_1 \mu_2} \log p(\mu \mid D) & \frac{\partial^2}{\partial \mu_2^2} \log p(\mu \mid D) \end{array} \right] \Bigg|_{\mu=\mu^\dagger} \right)^{-1}.$$

The log posterior distribution is

$$\log p(\mu \mid D) = \log \mathcal{N}\left(\mu_0, \sigma_0^2 I_2\right) + \sum_{t=1}^{n} \log \mathscr{P}\mathcal{N}(D_t; \mu, I_2) + C,$$

where C is a term that does not depend on μ, and the calculations to compute the Hessian are given in Hernandez-Stumpfhauser (2012). This Laplace refinement to the variational approximation greatly improves the quality of the original approximation, as is shown in Hernandez-Stumpfhauser (2012) by comparing the variational approximation and the variational/Laplace approximation to the output of the Gibbs sampler. Similar results hold in the regression case: the Laplace refinement substantially improves the quality of the variational approximation.

4. Comparing model selection via Gibbs, variational, and variational/Laplace

For a general Bayesian estimation problem, the deviance is defined as $\Delta(D, \omega) = -2 \ln p(D \mid \omega)$ where D are the data, ω are the unknown parameters and $p(D \mid \omega)$ is the likelihood function (Gelman et al. 2004, p. 179–184). The expected deviance $\mathrm{E}[\Delta(D, \omega) \mid D]$ is a measure of how well the model fits and it can be estimated by the posterior mean deviance $\overline{\Delta(D)} = B^{-1} \sum_{b=1}^{B} \Delta(D, \omega^{(b)})$, where $\{\omega^{(b)}\}_{b=1}^{B}$ are

random draws from the posterior distribution. The difference between the posterior mean deviance and the deviance at the posterior mean, estimated as

$$p_\Delta = \overline{\Delta(D)} - \Delta(D, \bar{\omega})$$

where $\bar{\omega} = B^{-1} \sum_{b=1}^{B} \omega^{(b)}$, is often interpreted as a measure of the effective number of parameters of a Bayesian model. More generally, p_Δ can be thought of as the number of "unconstrained" parameters in the model, where a parameter counts as 1 if it is estimated without constraints or prior information, 0 if it is fully constrained or if all the information about the parameter comes from the prior distribution, or an intermediate value if both the data and prior distributions are contributing.

We used the Deviance Information Criterion,

$$\text{DIC} = 2\overline{\Delta(D)} - \Delta(D, \bar{\omega}) = \overline{\Delta(D)} + p_\Delta,$$

to compare different model specifications for the departure time data. The DIC can be interpreted as a measure of goodness-of-fit (the estimated expected deviance) plus a penalty for model complexity in the form of the total number of effective parameters. Lower values of DIC correspond to more preferable tradeoffs between fit and model complexity.

We evaluated a large number of different model specifications for the mean of the projected normal distribution, including fixed and random effects for the states, waves and modes as well as for interactions between these factors. As an example of the type of models compared, the following is the full hierarchical specification for a model with mode as fixed effect and state and wave as random effects:

$$\begin{aligned}
D_{ijkt} &\sim \mathscr{PN}(\mu_{ijk}, I_2) \\
\mu_{ijk} &= \mu + m_k + s_i + w_j \\
\mu &\sim \mathscr{N}(0, 10^6 I_2) \\
m_k &\sim \mathscr{N}(0, 10^6 I_2) \\
s_i &\sim \mathscr{N}(0, \sigma_s^2 I_2) \\
w_j &\sim \mathscr{N}(0, \sigma_w^2 I_2) \\
\sigma_s^2 &\sim \mathscr{IG}(0.001, 0.001) \\
\sigma_w^2 &\sim \mathscr{IG}(0.001, 0.001).
\end{aligned}$$

In this specification, μ, m_k have vague priors while those of s_i, w_j are determined by their variance parameters, which follow pre-specified inverse gamma hyper-priors. This hierarchical set-up is similar to the usual Bayesian normal regression model.

We computed DIC and p_Δ values using Gibbs sampling, variational and variational/Laplace. Table 1 shows the Gibbs DIC values for different models applied to the departure time data. In Table 1, the models containing all three factors (mode, state, wave) consistently achieve lower DIC values than the models that excluded any of those factors. While not shown here, models with mode as random effect performed worse than models with mode as fixed effect. In contrast, very similar DIC values were obtained with the state and wave treated as either fixed or random. When we investigated models with interactions between the three factors, those with state-wave interactions scored better than any other arrangement of two-way interactions. Among the various models considered, DIC leads to selection of

$$\mu_{ijk} = \mu + m_k + sw_{ij},$$

with m_k a fixed mode effect and sw_{ij} a random interaction effect between the state and wave, with 99 total levels. Note that there are $6 \times 18 = 108$ possible state-wave combinations, so that there are nine state-wave combinations without observations where a mode-only model was applied. This was the final model used for purposes of small area estimation.

The effective number of parameters p_Δ for each model, computed via Gibbs sampling, are shown in Table 2. In interpreting these values, it should be noted that one level of a factor is represented by two parameters. Hence, in a model with only a mode effect there are eight parameters: two for the overall mean and six more for the three remaining free mode levels. The model with only a mode effect has p_Δ values (in the first row of Table 1) very close to eight. The final selected model has $p_\Delta = 191.5$.

We now turn to a comparison of the computation of DIC and p_Δ using Gibbs, variational and variational/Laplace. All three methods yield essentially identical posterior means $\bar{\omega}$, so $\Delta(D, \bar{\omega})$ is also essentially identical across methods. The differences in DIC across methods, displayed in Table 1, and differences in p_Δ across methods, displayed in Table 2, therefore come from differences in the posterior mean deviance $\overline{\Delta(D)}$ across methods. As can be seen from the two tables, the variational approximation without Laplace refinement significantly underestimates the posterior mean deviance, resulting in large negative differences in both DIC and p_Δ values. By contrast, Gibbs and variational/Laplace yield nearly identical estimates of the posterior mean deviance, hence virtually identical DIC and p_Δ values. For the tabled results, iterating the variational/Laplace approximation to convergence is about 15 times faster than 5000 iterates of Gibbs sampling. For purposes of model selection, therefore, the variational/Laplace approximation performs extremely well in this example.

Table 1: DIC values from Gibbs sampler for ten different projected normal model specifications, along with comparisons to DIC computed via other methods: Variational DIC minus Gibbs DIC and Variational/Laplace DIC minus Gibbs DIC.

Fixed Effects	Random Effects	Gibbs DIC	Variational − Gibbs DIC	Variational/Laplace − Gibbs DIC
mode		2642714.6	−2.7	−0.5
mode; wave		2631925.2	−4.3	0.1
mode	wave	2631925.9	−5.1	0.0
mode; state		2626382.7	−18.1	0.7
mode	state	2626383.6	−20.0	0.1
mode; wave; state		2616177.1	−23.4	0.2
mode; state	wave	2616177.2	−23.5	−1.7
mode; wave	state	2616175.4	−21.5	1.3
mode	state; wave	2616176.3	−22.9	−0.4
mode	state×wave	2613338.4	−105.9	−0.4

Table 2: Effective number of parameters p_Δ values from Gibbs sampler for ten different projected normal model specifications, along with comparisons to effective number of parameters computed via other methods: Variational p_Δ minus Gibbs p_Δ and Variational/Laplace p_Δ minus Gibbs p_Δ.

Fixed Effects	Random Effects	Gibbs p_Δ	Variational − Gibbs p_Δ	Variational/Laplace − Gibbs p_Δ
mode		8.3	−1.3	−0.2
mode; wave		17.7	−2.1	0.1
mode	wave	18.0	−2.5	0.0
mode; state		41.4	−9.0	0.4
mode	state	41.8	−9.9	0.1
mode; wave; state		52.5	−11.7	0.1
mode; state	wave	52.5	−11.7	−0.8
mode; wave	state	51.5	−10.7	0.7
mode	state;wave	52.0	−11.4	−0.2
mode	state×wave	191.5	−53.4	−0.9

5. Discussion

In this paper, we have briefly described an important small area estimation problem in which a hierarchical linear model is embedded in a nonlinear, projected normal model. A massive data set is considered, for which MCMC is feasible but slow. A large number of models are compared. Though a mean field variational approximation is not very accurate in this problem, it can be refined substantially by using a Laplace approximation, and the resulting variational/Laplace approximation is both accurate and extremely fast to compute. In particular, model selection results are virtually indistinguishable between the MCMC and the variational/Laplace approaches. While these results are limited to the particular problem under consideration, they do suggest that there is considerable promise for variational/Laplace approximations in model selection and inference in small area estimation problems.

REFERENCES

BISHOP, C. M., (2006). *Pattern Recognition and Machine Learning*. Springer.

FAY, R. E., HERRIOT, R. A., (1979). Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association 74*, 269–277.

GELMAN, A., CARLIN, J. B., STERN, H. S., RUBIN, D. B., (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

HERNANDEZ-STUMPFHAUSER, D., (2012). *Topics in Design-Based and Bayesian Inference for Surveys*. Ph. D. thesis, Colorado State University.

NEAL, R. M., (2003). Slice Sampling. *The Annals of Statistics 31*, 705–741.

NUÑEZ-ANTONIO, G., GUTIÉRREZ-PEÑA, E., (2005). A Bayesian Analysis of Directional Data Using the Projected Normal Distribution. *Journal of Applied Statistics 32*(10), 995–1001.

NUÑEZ-ANTONIO, G., GUTIÉRREZ-PEÑA, ESCALERA, E. G., (2011). A Bayesian Regression Model for Circular Data Based on the Projected Normal Distribution. *Statistical Modeling 11*, 185–201.

ORMEROD, J. T., WAND, M. P., (2010). Explaining Variational Approximations. *The American Statistician 64*, 140–153.

PRESNELL, B., MORRISON, S. P., LITTELL, R. C., (1998). Projected Multivariate Linear Models for Directional Data. *Journal of the American Statistical Association 93*(443), 1068–1077.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 105–132

DEVELOPMENT OF SMALL AREA ESTIMATION IN OFFICIAL STATISTICS[1]

Jan Kordos[2]

ABSTRACT

The author begins with a general assessment of the mission of the National Statistics Institutes (NSIs), main producers of official statistics, which are obliged to deliver high quality statistical information on the state and evolution of the population, the economy, the society and the environment. These statistical results must be based on scientific principles and methods. They must be made available to the public, politics, economy and research for decision-making and information purposes.
Next, before discussing general issues of small area estimation (SAE) in official statistics, the author reminds: the methods of sampling surveys, data collection, estimation procedures, and data quality assessment used for official statistics. Statistical information is published in different breakdowns with stable or even decreasing budget while being legally bound to control the response burden.
Special attention is paid, from a practitioner point of view, to synthetic development of small area estimation in official statistics, beginning with international seminars and conferences devoted to SAE procedures and methods (starting with the Canadian symposium, 1985, and the Warsaw conference, 1992, to the Poznan conference, Poland, 2014), and some international projects (EURAREA, SAMPLE, BIAS, AMELI, ESSnet). Next, some aspects of development of SAE in official statistics are discussed. At the end some conclusions regarding quality of SAE procedures are considered.

Key words: small area estimation, official statistics, sampling survey, direct estimation, indirect estimation, empirical Bayes estimator; hierarchical Bayes estimator; data quality.

1. Introduction

National Statistics Institutes (NSIs) are the most important statistical information providers for official statistics. Their mission is to produce high

[1] This is an updated and extended version of the first part of the paper entitled "Small Area Estimation in Official Statistics and Statistical Thinking" presented at the International Conference on Small Area Estimation 2014, held in Poznan, Poland, 3-5 September 2014.

[2] Central Statistical Office of Poland and Warsaw Management Academy.

quality statistical information on the state and evolution of the population, the economy, the society and the environment. These statistical results must be based on scientific principles and methods. They must be made available to the public, politics, economy and research for decision-making and information purposes. One important challenge that NSIs have to face is the growing users' demand with stable or even decreasing budget while being legally bound to control the response burden. The use of more and more efficient statistical methods is a way to take up this challenge. To collect, estimate, process and publish statistical information NSIs use different methods and procedures, but special emphasis is paid to sampling surveys, taking into account basic needs, cost and respondent burden. For this reason, issues connected with sampling surveys in official statistics from a practitioner point of view are considered first here, using different approaches, methods, and variety of data, mainly sampling data, censuses and registers (Brakel & Bethlehem, 2008; Little, 2004. 2012).

2. Sampling surveys in official statistics and issues of SAE methods

First, the author would like to remind that the purpose of sampling surveys is to obtain statistical information about a finite population by: a) selecting a probability sample from this population, b) obtaining or measuring the required information about the units in this sample, and c) estimating finite population parameters such as means, totals, ratios, etc., and assessing their variances (Brakel & Bethlehem, 2008). The statistical inference in this setting can be: (i) *design-based*, (ii) *model-assisted* or (iii) *model-based*. In the design-based and model-assisted approach, the statistical inference is based on the stochastic structure induced by the sampling design. Parameter and variance estimators are derived under the concept of repeatedly drawing samples from a finite population according to the same sampling design, while statistical modelling plays a minor role. This is the traditional approach of survey sampling theory, followed by authors like Hansen et al. (1953), Kish (1965), Cochran (1977), Yates (1981) and Särndal et al. (1992).

In the model-based context, the probability structure of the sampling design plays a less pronounced role, since the inference is based on the probability structure of an assumed statistical model. This is the position taken by authors like Gosh and Meeden (1997), Gosh &Rao (1994), Rao (1999), Valliant et al. (2000), Rao (2003), Pfeffermann (2002, 2013) and Jang & Lahiri (2006). An overview of the different modes of inference in survey sampling is given by Little (2004).

Design-based and model-assisted estimators refer to a class of estimators that expand or weight the observations in the sample with the so-called survey weights. Survey weights are derived from the sampling design and available auxiliary information about the target population. Functions of the expanded observations in the sample are used as (approximately) design-unbiased

estimators for the unknown population parameters of interest. The associate inferences are based on the probability distribution induced by the sampling design with the population values held fixed.

A well-known design-based estimator is the π-estimator or Horvitz-Thompson estimator, developed by Narain (1951), and Horvitz and Thompson (1952) for unequal probability sampling from finite populations without replacement. The observations are weighted with the inverse of the inclusion probability, also called design-weights. This estimator is design-unbiased, since the expectation of the estimator with respect to the probability distribution induced by the sampling design is equal to the true but unknown population value.

The precision of the Horvitz-Thompson estimator can be improved by making advantage of available auxiliary information about the target population (Wywiał, 2000). In the model-assisted approach developed by Särndal et al. (1992) this estimator is derived from a linear regression model that specifies the relationship between the values of a certain target parameter and a set of auxiliary variables for which the totals in the finite target population are known. Based on the assumed relationship between the target variable and the auxiliary variables, a generalized regression estimator can be derived of which most well-known estimators are special cases. After this estimator is derived, it is judged by its design-based properties, such as design expectation and design variance.

Most NSIs surveys are designed to provide statistically reliable estimates at national or high-level geographies. So when statistics are required for more detailed geographical areas or small subgroups of the population, the sample sizes just are not big enough to make reliable estimates. Increasing the size of samples would be prohibitively expensive – instead, estimation methods have been developed that combine data from administrative, census and survey sources to produce estimates for small areas or domains. There are many statistical techniques covered by small area estimation, a frequently used approach is a model-based one, where local area outcomes are estimated from the regression between survey data and auxiliary data from census and administrative data sources.

The great importance of SAE stems from the fact that many new programs, such as fund allocation for needed areas, new educational or health programs and environmental planning rely heavily on these estimates. SAE techniques are also used in many countries to test and adjust the counts obtained from censuses that use administrative records and for post-enumeration surveys after the population censuses for quality assessment. SAE is researched and applied so broadly because of its usefulness to researchers who wish to learn about the research carried out in SAE and to practitioners who might be interested in applying the new methods.

The problem of SAE is twofold. First, the fundamental question is how to produce reliable estimates of characteristics of interest (means, counts, quantiles, etc.) for small areas or domains, based on very small samples taken from these areas. The second related question is how to assess the estimation error. Budget

and other constraints usually prevent the allocation of sufficiently large samples to each of the small areas. Also, it is often the case that domains of interest are only specified after the survey has already been designed and carried out. Having only a small sample (and possibly an empty sample) in a given area, the only possible solution to the estimation problem is to *borrow information* from other related data sets.

As it has been mentioned, and from theoretical point of view, SAE methods can be divided broadly into "*design-based*" and "*model-based*" methods. The latter methods use either the frequentist approach or the full Bayesian methodology, and in some cases combine the two, known in the SAE literature as "*empirical Bayes".* Design-based methods often use a model for the construction of the estimators (known as "*model assisted*"), but the bias, variance and other properties of the estimators are evaluated under the randomization (design-based) distribution. The randomization distribution of an estimator is the distribution over all possible samples that could be selected from the target population of interest under the sampling design used to select the sample, with the population measurements considered as fixed values (parameters). Model-based methods, on the other hand, usually conditioned on the selected sample, and the inference is with respect to the underlying model. A common feature to design- and model-based SAE is the use of auxiliary covariate information, as obtained from large surveys and/or administrative records such as censuses and registers. Some estimators only require knowledge of the covariates for the sampled units and the true area means of these covariates. Other estimators require knowledge of the covariates for every unit in the population. The use of auxiliary information for SAE is vital because with the small sample sizes often encountered in practice, even the most elaborated model can be of little help if it does not involve a set of covariates with good predictive power for the small area quantities of interest.

It is now generally accepted that the indirect estimates should be based on explicit models that provide links to related areas through the use of supplementary data such as census counts or administrative records. See, for example, Ghosh and Rao (1994), Rao (1999), Rao (2003), Pfeffermann (2002, 2013), and Jiang and Lahiri (2006) for more discussion on model-based small area methods.

Thus, the model-based estimates are obtained to improve the direct design-based estimates in terms of precision and reliability, *i.e.*, smaller coefficients of variation (CVs). Supplementary data are vital for improving quality of small area statistics. These data are used to construct predictor variables for use in a statistical model that can be used to predict the estimate of interest for small areas. The effectiveness of small area estimation depends initially on the availability of good predictor variables that are uniformly measured over the total area. It next depends on the choice of a good prediction model. Effective use of small area estimation methods further depends on a careful, thorough evaluation of the quality of the model. Finally, when small area estimates are produced, they should be accompanied by valid measures of their precision. Now, there is a wide

range of different, often complex models that can be used, depending on the nature of the measurement of the small area estimates and on the auxiliary data available. One key distinction in model construction is between situation where the auxiliary data are available for the individual units in the population and those where they are available at aggregate level for each small area. In the former case, the data can be used in unit level models. Another feature involved in the choice of a model is whether the model borrows strength across sectional or over time, or both. There are also now a number of different approaches, such as empirical best linear prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes (HB) that can be used to estimate the models and the variability of the model dependent small area estimates (Choudry et al., 1989, 2012; Data. 2009; Datta et al.,1999, 2012; Gosh & Meeden, 1997, 1999; Kubacki, 2004; Lehtonen et al., 2003, 2005, 2009; Molina et al., 2009; Moura et al., 2002; Pfeffermann,1999, 2013; Pfeffermann & Tiller, 2006; Pratesi &Salvati, 2008; Rao, 2003, 2011; You & Dick, 2004). Moreover, complex procedures that would have been extremely difficult to apply a few years ago can now be implemented fairly straightforwardly, taking advantage of the continuing increases in computing power and the latest developments in software.

Thus, there are two broad classifications for small area models: ***area level models*** and ***unit level models***:

- ***Area level models*** that relate small area means and totals to area-specific auxiliary variables,
- ***Unit level models*** that relate the unit values of the dependent variable to unit specific auxiliary variables.

Among the ***area level models,*** the ***Fay-Herriot model*** (Fay and Herriot, 1979) *is a basic and widely used area level model in practice to obtain reliable model-based estimates for small areas.* The Fay-Herriot model basically has two components, namely, a sampling model for the direct estimates and a linking model for the parameters of interest. The sampling model involves the direct survey estimate and the corresponding sampling variance. The Fay- Herriot model assumes that the sampling variance is known in the model. Typically, a smoothed estimator of the sampling variance is obtained and then treated as known in the model. Wang and Fuller (2003), You and Chapman (2006), Gonzalez-Manteiga, et al. (2010), considered the situation where the sampling variances are unknown and modelled separately by direct estimators.

The linking model relates the parameter of interest to a regression model with area-specific random effects. In the Fay-Herriot model, the area random effects are usually assumed to be independent and identically distributed normal random variables to capture geographically unstructured variations among areas. However, in some small area applications, particularly in public health estimation problems, geographical variation of a disease is a subject of interest, and estimation of overall spatial pattern of risk and borrowing strength across regions to reduce variances of final estimates are both important. Thus, it may be more

reasonable to construct spatial models on the area-specific random effects to capture the spatial dependence among them. The spatial models are generally used in health related small area estimation, and various spatial models have been proposed for small area estimation [(*e.g.* Ghosh et al., 1999; Moura et al., 2002; Pratesi and Salvati (2008), Singh et al., (1994) and Molina et al., (2009)]. Best et al., (2005) provided a comprehensive review on spatial models for disease mapping. Rao (2003) also discussed several spatial small area models.

The unit model originates with Battese, Harter and Fuller (1988). They used the nested error regression model to estimate county crop areas using sample survey data in conjunction with satellite information. Prasad and Rao (1990, 1999) were first to include the survey weights in the unit level model: they labelled their estimator as a pseudo-EBLUP estimator of the small area mean. Prasad and Rao (1999) also provided based expressions for the MSE of their estimator when it included the estimated variance components. You and Rao (2002) proposed an estimator of β that ensures self-benchmarking of the small area estimates to the corresponding direct estimator.

Thus, an *indirect estimator* uses values of the variable of interest from a domain and/or time period other than the domain and time period of interest. Three types of indirect estimators can be identified:

- A *domain indirect estimator* uses values of the variable of interest from another domain but not from another time period.
- A *time indirect estimator* uses values of the variable of interest from another time period but not from another domain.
- An estimator that is both *domain and time indirect* uses values of the variable of interest from another domain and another time period.

Individual level models work in two stages using regression modelling. Firstly, the survey data are used to predict the probability of the characteristic of interest based on the attributes of the individuals in the survey (such as gender, age and marital status). The aggregate levels of a cross tabulation of these individual characteristics for each local area are obtained, usually from the census, and the coefficients from the regression model are applied to those small area covariate values so as to calculate the expected value of the target outcome variable conditional on the area's characteristics. The steps are relatively straightforward:

- Ensure that the predictor variables are available in both survey data and for small areas
- Fit a regression model to survey data to predict the probability of chosen outcome
- Use Wald tests to consider dropping non-significant variables.
- Extract parameter estimates and apply to small area data.

In short, SAE is a collection of different methods:

- *Synthetic methods* (often implicit assumptions on the nature of relationship; a simple case: rates at NUTS2 level = rates at NUTS4 level).
- *Composite methods* (linear combination of synthetic methods and direct estimators to balance bias and variance).
- *Estimator based on linear mixed models* (EBLUP, EB, HB).
- *Non-linear models* (e.g. logistic models for binary responses) with spatial and/or temporal correlation structures among random effects.
- *Semi-parametric models.*

Potentially more serious, with respect to *accuracy and quality*, are ***non-sampling errors*** **s**uch as ***coverage errors, measurement errors*** and ***response bias***. Most censuses miss some people, or count some people twice, and it has been repeatedly shown that those miscounted are generally not typical of the population as a whole. Census or sample survey estimates may therefore be biased against certain subgroups of the population. If these subgroups tend to be geographically clustered, this can have a serious impact on estimates for some small areas. Response bias arises if many respondents systematically misunderstand a census or a survey question or are unable or unwilling to give correct answer. Both small area and large area estimates would be affected by such errors (Bethlehem, 1988; Bethlehem et al. 1985; Brackstone, 1999, Eurostat, 2007; Holt et al, 1991; Kalton, 2002; Kalton & Kasprzy, 1986; Kordos, 2005; Longford, 2005; Rao, 2011; Trewin, 2002).

3. Use of administrative data in official statistics

NSIs around the world are coming under increasing pressure to improve the efficiency of the statistical production process, and particularly to make savings in costs and staff resources. At the same time, there are growing political demands to reduce the burden placed on the respondents to statistical surveys. Given these pressures, statisticians are increasingly being forced to consider alternatives to the traditional survey approach as a way of gathering data. Perhaps the most obvious answer is to see if usable data already exist elsewhere. Many nonstatistical organisations collect data in various forms, and although these data are rarely direct substitutes for those collected via statistical surveys, they often offer possibilities, sometimes through the combination of multiple sources, to replace, fully or partially, direct statistical data collection. The degree of the use of administrative sources in the statistical production process varies considerably from country to country, from those that have developed fully functioning register-based statistical systems, to those that are just starting to consider this approach. A significant contribution in this field is publication issued in 2011 by

the United Nations Economic Commission for Europe[3], entitled "***Using Administrative and Secondary Sources for Official Statistic, A Handbook of Principles and Practices***". These trends make model-based procedures more and more attractive and relevant for NSIs to apply in the production of official statistics (Chambers et al., 2006).

Administrative datasets are typically very large, covering samples of individuals and time periods not normally financially or logistically achievable through survey or even census methodologies. Alongside cost savings, the scope of administrative data is often cited as its main advantage for research purposes, though coverage is recognized to be imperfect. The lack of control the researcher has during the data collection stage and how this affects its quality, and therefore what can be done with the data, are the main problems for administrative data. More general concern has also been voiced about the lack of well-established theory and methodologies to guide the use of administrative data in social science research.

Potential auxiliary data should be evaluated for their relationship to the variable(s) of interest, both theoretically and statistically as well as the accuracy and reliability with which they have been collected. The theoretical relationship should emanate from tested social or economic theories. A careful examination should be made to understand any major differences between the auxiliary data and the variables of interest.

Consideration should be given to the purpose for which the data were initially collected, how it was processed and edited, what conceptual definitions were used and what the scope of the auxiliary data holdings is. This will allow appropriate auxiliary information to be chosen to improve the model, and in explaining to users what factors are driving the small area estimates and help pinpoint potential sources of error.

Although auxiliary information was originally used in the design and estimation procedure of a survey to decrease the sampling variance of estimators, nowadays it is an important tool to decrease the bias due to selective non-response. Estimators using auxiliary information are generally more robust against selective non-response than estimators that do not use auxiliary information (Bethlehem, 1988; Särndal et al., 1987, 2005; Thomsen et al., 1998).

Common concern around the use of detailed administrative data at the small area level includes risks around confidentiality, anonymity and disclosure and this may lead to data controllers refusing to release the data or making it available within very controlled environments. An important consideration therefore for the release or publication of administrative data at individual or aggregate small area level is that the identity of individuals is protected. The assessment of disclosure risk is a complex process. Generally, the more detail the data has and the higher

[3] http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf.

the proportion of the population of interest that is captured in the data, the higher the risk.

There are various ways in which extracts of administrative data can be linked with other data sources to create more comprehensive and powerful datasets for analysis (both in terms of cases and variables).

Examples of available administrative data: i) population, ii) building and dwellings, iii) taxes, iv) business registers. Uses of administrative data are especially useful for:

1) improving survey results (sampling frame for persons and business surveys; auxiliary variables for calibration);
2) reducing the respondent burden: directly (some questions are skipped); indirectly (gain efficiency the estimators).

Administrative data holds great research potential for SAE (and other) research in all national contexts although the research availability and use of such data varies significantly between countries. There is also a problem how to use available BIG Data or other approaches for SAE in official statistics.

4. The international conferences and research projects towards application of SAE methods in official statistics

There have been different kind of conferences, seminars, and research projects devoted to exchange of ideas, experiences and achievements related to application of SAE in official statistics. First, some international conferences and next selected international research projects devoted to applications of SAE methods are briefly presented.

4.1. The International Conferences to apply SAE methods in official statistics

The results of the first attempts of applications of SAE methods in official statistics were presented at the symposium held ***in Ottawa in 1985 and published in Platek et al. (1987)***.

This publication had significant impact on academic statisticians and research statisticians in NSIs, and specially on countries in transition in Central and Eastern Europe, which organized international conferences held *in Poland in 1992 (Warsaw Conference in 1992: Kalton et al., 1993), and Latvia in 1999 (Riga conference: Riga, 1999).*

Starting from 2005, a new series of SAE Conferences have taken place in: *Finland, (Jyvaskyla, 2005); Italy, (Pisa, 2007); Spain, (Elche, 2009); Germany, (Trier, 2011); Thailand, (Bangkok, 2013); Poland, (Poznan, 2014), Chile, (Santiago de Chile, 2015).*

SAE conferences are aimed at providing a platform for discussion and exchange of ideas about current developments in small area estimation research in different fields. The conferences address - in a good balance with theoretical and

methodological development in small area estimation and related fields, and in practical application - of SAE methods, including their potential uses in various research areas in official statistics. The need to regulate and promote the continuity of SAE conferences required the creation a working group with an acronym: EWORSAE – the European Working Group on Small Area Estimation[4] – aimed to build and maintain a network of researchers and statisticians to foster collaborative work and to increase cooperation between Statistical Offices and the research community within the field of SAE and related topics. Although the working group is basically European, it is open to all people worldwide working in small area estimation[5].

4.2. The International Projects for SAE implementations in official statistics

Before presenting some international projects for SAE methods applications in official statistics, it seems reasonable to begin with a program started in the USA over 20 years ago.

SAIPE – an acronym for ***Small Area Income and Poverty Estimates***[6]. The U.S. Census Bureau's program started at the beginning of 1990s and has provided annual estimates of income and poverty statistics for all states, counties, and school districts. The main objective of this program is to provide estimates of income and poverty for the administration of federal programs and the allocation of federal funds to local jurisdictions. In addition to these federal programs, state and local programs use the income and poverty estimates for distributing funds and managing programs. SAIPE revises and improves methodology as time and resources allow. The details of the methodology differ slightly from year to year. The most significant change was between 2004 and 2005, when SAIPE began using data from the *American Community Survey*, rather than from the *Annual Social and Economic Supplement to the Current Population Survey.*

Some impact on applications of SAE procedures in official statistics has had the following international projects sponsored by the European Union:

EURAREA; SAMPLE; BIAS; AMELI; ESSnet

4.2.1. The EURAREA project investigated methods for small area estimation and their application in official statistics. It was funded by Eurostat under the Fifth Framework (FP5) Programme of the European Union and was carried out by a consortium of NSIs, universities and research consultancies from across the European Union (United Kingdom, Spain, Italy, Sweden, Norway, Finland and Poland). The project was co-ordinated by the UK Office for National Statistics. It ran from January 2001 until June 2004 and was signed off by

[4] On initiative of Spanish Statisticians.
[5] http://sae.wzr.pl/.
[6] http://beta.census.gov/did/www/saipe/about/index.html.

Eurostat in February 2005. The aim of this project was to evaluate the effectiveness of standard estimation techniques for small areas (synthetic estimators, GREGs and composite estimators). The studies carried out until 2004 were based on sampling designs with equal selection probabilities. In order to undertake this project it was necessary to study the existing theory as well as to develop new theories that make it easier to obtain estimation techniques and their mean squared error when other sampling plans are used that are more similar to those applied in official statistics in the real world. Finally, all the theory developed has been implemented in a SAS IT application whose use has been widely documented so that any user is able to apply the programme to his/her own data. The links below provide further information about the project, its aims, objectives and conclusions[7].

The research outputs from the project are available in the download section: these include the final project reference volume and macro language programs written in SAS. The project reference volume contains reviews of existing theory in small area estimation, an assessment of the "standard" estimators and the results of the innovative work undertaken within the project. The program codes for the procedures investigated are provided so that the results can be implemented by other NSIs and statisticians. The program code has been written in SAS macro language or SAS macros or routines that can be called in SAS. Some results are also presented in EURAREA (2004), Heady et al. (2001, 2004) and Chambers et al. (2006).

4.2.2. SAMPLE: Small Area Methods for Poverty and Living Condition Estimates

The Project was supported by the European Commission (FP7-SSH-2007-1). The aim of SAMPLE project was to identify and develop new indicators and models for inequality and poverty with attention to social exclusion and deprivation, as well as to develop, implement models, measures and procedures for *small area estimation* of the traditional and new indicators and models[8]. This goal was achieved with the help of the local administrative databases. Local government agencies often had huge amount of administrative data to monitor some of the actions which witness situations of social exclusion and deprivation (social security claims for unemployment and eligibility for benefits from any of the programs Social Security administers) of households and citizens. SAMPLE utilised widely used indicators on monetary and non-monetary poverty. Moreover, in collaboration with stakeholders working with the poor, the project developed new poverty indicators that meet local needs. The results of the SAMPLE project will help local authorities and stakeholders to plan and implement their poverty-reduction policies. In fact, more than two thirds of

[7] http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html.

[8] http://www.bing.com/search?ei=UTF-8&pc=AV01&q=http%3A%2F%2Fwww.sample-project.eu%2F&FROM=AVASDF.

stakeholders surveyed said that these local indicators would prove very useful in the planning of social policies. The final goal of the project is to provide a dashboard of reliable indicators of poverty and deprivation defined at NUTS3, NUTS4 level, useful for Local Government Agencies. In the project, the EU-SILC[9] sample will be enlarged at NUTS4[10].

The Project was coordinated by Prof. Monica Pratesi, Italy. Consortium of the project: Prof. Achille Lemmi, Italy; Dr Nikos Tzavidis, UK; Dr Isabel Molina, Spain ; Prof. Domingo Morales, Spain; Prof. Tomasz Panek, Poland; Dr Paolo Prosperini, Dr Claudio Rognini, Dr Moreno Toigo, Italy.

4.2.3. BIAS Project

The BIAS project is an acronym for "*Bayesian methods for combining multiple Individual and Aggregate data Sources in observational studies*" project. The first edition of the project named BIAS I was funded between April 2005 and June 2008 under the first phase of node commissioning. The second edition named BIAS II was funded by the second commissioning phase from July 2008 to June 2011. Description of BIAS I and BIAS II projects is based on information from the web page: www.bias-project.org.uk.

The aims of the project were: a) to develop a set of statistical frameworks for combining data from multiple sources, b) to improve the capacity of social science methods to handle the intricacies of observational data. In this project Bayesian hierarchical models are used as the basic building blocks for these developments. These offer a natural tool for linking together many different sub-models and data sources. The BIAS I research programme consisted of three methodological components: a) multiple bias modelling for observational studies, b) combining individual and aggregate level data, c) small area estimation.

The last one was especially devoted to small area estimation methodology and was being carried out in collaboration with ONS. The basic methodological problem was to estimate the value of a given indicator (e.g. income, crime rate, unemployment) for every small area, using data on the indicator from individual-level surveys in a partial sample of areas, plus relevant area-level covariates available for all areas from census and administrative sources, for example.

4.2.4. The AMELI Project

The project AMELI (*Advanced Methodology for Laeken Indicators*) was a trial to satisfy expectations of the need for effective, high-quality, robust, timely and reliable statistics and indicators related to the social cohesion. It started in April 2008 and ended in March 2011. The main target of the project was to review the state-of-the-art of the existing indicators monitoring the multidimensional phenomena of poverty and social exclusion - the Laeken indicators including their relation to social cohesion. Special emphasis was put on

[9] EU-SILC – an acronym for: European Union Statistics on Income and Living Conditions.
[10] NUTS – an acronym for: Nomenclature of Units for Territorial Statistics.

methodological aspects of indicators and especially on their impact on policy making. This included quality aspects as well as mathematical and statistical properties within a framework of a complex survey in the context of practical needs and peculiarities. The official website of the project is: http://ameli:surveystatistics:net/.

The coordinator of the Project was Prof. Ralf Münnich, University of Trier, Germany. Consortium consisted of: Federal Statistical Office of Germany, Swiss Federal Statistical Office, Statistics Austria, Statistics Finland, University of Helsinki, Vienna University of Technology, Statistical Office of Slovenia, Statistics Estonia.

4.2.5. The ESSnet Project for SAE

The project lasted 27 months (December 2009 to March 2012). The coordinator of the Project was: Stefano Falorsi, ISTAT, Italy. Co-partners: INSEE, France; DESTATIS, Germany; CBS, Netherlands; SSB, Norway; GUS, Poland; INE, Spain; ONS, United Kingdom; SFSO, Switzerland.

The general objective of the project was to develop a framework enabling the production of small area estimates for ESS social surveys.

The specific objectives were to: a) complete (thestate of the art level) the EURAREA project, b) update the documents available on small area estimation, c) describe the current application in UE NSIs and non-UE NSIs, d) create a common knowledge on application of small area estimation methods; e) review and develop suitable criteria to assess the quality of SAE methods for the choice of proper model and the evaluation of MSE; f) make available software tools for SAE to the ESS; g) foster knowledge transfer by the development of case studies and associated recommendations on representative problems in small area estimation in the ESS; h) provide practical guidelines in ESS social surveys context; i) transfer knowledge and know-how to non-participating NSIs and disseminate results.

Results of the project (lessons learned):

1) The work done and the outcomes produced by the project are strategic for increasing the capability within ESS to produce official statistics by SAE techniques.
2) The upload of all outcomes within the EU-cross-portal is very useful for disseminating scientific and applicative results of the ESSnet.
3) It should be useful to try to develop the regular exchange of information about SAE methods and applications among NSIs giving impulse to the use of forum within the website.
4) The course was very useful for involving the non-participating NSIs and transferring the results of the project within the ESS. It was also useful in order to map the real needs of non-participating countries.
5) The different presentations in scientific workshops and conferences were important to disseminate the knowledge of the outcomes of the project.

6) The survey on the use of methods and available tools within the NSIs of ESS and other NSIs has been very useful to map capabilities and application needs. This survey should be updated regularly and published into the website.

The results of the ESSnet project for SAE are strategic for increasing the capability to produce official statistics.

5. Discussion

Small area estimation methods have been developing significantly over the last 30 years and used partially in official statistics. Before small area estimates can be considered fully credible, carefully conducted evaluation studies are needed to check on the adequacy of the model being used. Sometimes model-dependent small area estimators turn out to be of superior quality to sample-based estimators, and this may make them seem attractive.

SAE techniques are becoming a matter of great interest for a variety of people, including statisticians, researchers and other university experts, and institutions, as NSIs, research institutes, governmental bodies, local authorities and private enterprises dealing with research methodology, empirical research and statistics production for regional areas and other population subgroups.

SAE methodologies have become a widely used method across various disciplines as a result of growing policy makers and researchers' demand for spatially detailed information alongside advances in small area data availability and computing power. Currently, despite the potential of these approaches and the growing demands placed upon them, there is little agreement within the academic and policy community as to which method(s) work best, whether different approaches are best suited to different local contexts, how best methods can be implemented and how best results can be validated. Experts from across each of these methodological strands and across a range of academic disciplines are included in the network so as to enable not only improvements in each separate approach but also overall methodological progress through the cross-pollination of ideas and skills.

Accuracy is generally considered to be a key measure of quality. Total survey error is a conceptual framework describing errors that can occur in a sample survey and the error properties. It may be used as a tool in the design of the survey, working with accuracy, other quality characteristics, and costs. Accuracy is often measured by the mean squared error (MSE) of the estimator. Error sources are considered one by one to estimate the uncertainty and also to obtain some indication of the importance of that source. The errors arise from: sampling, frame coverage, measurement, non-response, data processing, and model assumptions.

Therefore, indirect estimators are constructed that borrow strength from related areas, increasing the effective sample size and with it the estimation precision. These indirect estimators are based on either explicit or implicit models

providing a link between the small area in question and related areas through ancillary information. These auxiliary variables can be miscellaneous, cross-sectional as well as across time, for example information from neighbouring or next higher populations, data from a previous census or administrative records. Due to the growing demand for reliable small area statistics, small area estimation is becoming an important field in survey sampling.

Weighting is a statistical technique commonly used and applied in practice to compensate for nonresponse and coverage error. It is also used to make weighted sample estimates conform to known population external totals. In recent years a lot of theoretical work has been done in the area of weighting and there has been a rise in the use of these methods in many statistical surveys conducted by NSIs around the world.

In the last decade, calibration has been used to reduce both sampling error and nonresponse bias in surveys. In the presence of auxiliary variables with known population totals or with known values on the originally sampled units, the calibration procedure generates final weights for observations that, when applied to those auxiliary variables, yield their population totals or unbiased estimates of these totals, respectively. Unfortunately, in practice availability of such of auxiliary variables is rather not often.

The move to a more overt modelling approach means that government agencies need to recruit and train statisticians who are adept in modelling methods, as well as being familiar with survey sampling design. Survey sampling needs to be considered a part of mainstream statistics, in which Bayesian models that incorporate complex design features play a central role. A Bayesian philosophy would improve statistical output, and provide a common philosophy for statisticians and researchers in substantive disciplines such as economics and demography. A strong research program within government statistical agencies, including cooperative ties with statistics departments in academic institutions, would also foster examination and development of the viewpoints (Lehtonen et al. 2002, Lehtonen and Sarndal 2009).

5.1. Results of international conferences and projects

It is difficult to assess the impact of the international conferences and different projects on application of SAE methods in official statistics. General conclusion is that development and results of SAE methods in official statistics obtained so far from these conferences and the international projects have been mostly academic. Several projects aimed at development of SAE methodology such as *EURAREA, SAMPLE, BIAS, AMELI*, etc. are either completed or still ongoing at a country level. Next to these methodologies-oriented projects, quite few projects focused on estimating variables for social surveys undertaken by some NSIs. What is more, methodological know-how and techniques in SAE differ in NSIs. Some of NSIs have a great deal to offer in terms of expertise, links with academic experts and experience of implementation of these techniques while some others

are just at the empirical stage of practice. The first projects aiming at the development of SAE methodology did not highlight differences between European NSIs in the way they introduce the SAE methodology into the process of producing statistics. Only *ESSnet project for SAE* provides an overview of applications to social statistics for many European and some non-European NSIs. Furthermore, this project describes the research of the NSIs concerning SAE, which eventually will lead to a greater number of applications.

As it has been already stressed, the application of model-based estimation procedures in official statistics is limited. Several factors have been mentioned for the slow adoption of these methods. One is the fact that many NSIs are rather reserved in the application of model-based estimation procedures and generally rely on the more traditional design-based or model-assisted procedures for producing their official statistics. NSIs need to play safe in the production of official statistics and therefore do not want to relay on model assumptions, particularly if they are not verifiable (Chambers et al, 2006; Brakel & Betheyem, 2008; Eurarea, 2004; Little, 2004, 2012).

The availability of small area data has improved dramatically since the 1990s yet many spatial variables of interest – income, fear of crime, health-related behaviours, and so the list goes on – remain impossible to access at small area geographies in many national contexts. Within this context SAE methodologies have become increasingly demanded, increasingly used and increasingly refined. Yet the methodological landscape around SAE remains in need of attention in at least three key ways, according to Whitworth A. (edt)[11]. "Firstly, various alternative SAE methodologies have emerged and it is often unclear to some researchers what these alternative approaches are, how they relate to each other and how they compare in terms of their estimation performance. These methodological approaches can be classified broadly either as spatial microsimulation (which tend to be used by geographers predominantly) or statistical approaches (the use of which is dominated by statisticians). Secondly, despite recent advances in SAE methodologies there remain key methodological challenges and uncertainties to explore (e.g. how exactly each method can be best implemented in relation to weights, constraints, seeding, etc.) as well as innovative methodological advances to be brought together and extend (e.g. any role for agent-based modelling, estimating distributional functions or spatially varying interactions). Thirdly, the different methodological approaches to SAE in large part operate in parallel to one another without a clear understanding of the conceptual and methodological linkages between them. This is particularly true between the statistical and spatial microsimulation approaches and greater understanding of the linkages between methodologies within these two differing approaches could support important contributions to the effectiveness of current SAE best practice".

[11] http://eprints.ncrm.ac .uk/3210/1/sme_whitworth.pdf.

Nevertheless, SAE methods have been used in applications including *employment and unemployment statistics, health, poverty, agriculture, business, demography, census undercount, ecology, and education* (Datta et al.,1999, 2002; Dehnel, 2010; Dick, 1995; Drew et al., 1982; Elazar, 2004; Esteban et al., 2012; Gambino et al., 1998, 2000; Dehnel et al., 2004; Golata, 2004; Hidiroglou et al., 1985, 2007; Kordos, 1994, 2006; Kubacki, 2004; Molina et al., 2010; Paradysz, 1998; Paradysz & Dehnel, 2005; Schaible et al., 1994).

5.2. Differentiation in utilization of SAE methods by NSIs

As it has been stressed, NSIs are facing increasing demand for statistics below the level for which most large scale surveys have been designed. The survey methodologists are turning toward SAE techniques to satisfy the need for reliable estimates for small domains.

However, there are some common characteristics connected with applications of SAE procedures in official statistics. Usually such applications are prepared and implemented in cooperation with academic statisticians or subject-matter specialists and official statisticians. Very often there are still R&D approaches. It is impossible to discuss the differences by countries here, but the author confines himself to some issue connected with R&D in this field and quality aspects of the results. The author has found a number of very interesting publications in the Internet connected with applications of SAE methods in different fields and countries. Some of them include: Statistics Canada[12]; USA – Bureau of Census[13]; U.K- Office for National Statistics (ONS)[14] and Australian Bureau of Statistics[15]. The author would like to add the network, funded by the ESRC's National Centre for Research Methods (NCRM) Programme, which brings together experts in small area estimation techniques from the academic and policy (e.g. Office for National Statistics) communities in the UK and internationally in order to seek innovative ways to advance knowledge and understanding in SAE methodologies[16].

As it has already been stressed, it is impossible to discuss the differentiation of application of SAE procedures in different countries here, but the following issues will be considered: a) *Assessing the quality of small area estimates; b) Communicating quality to users.*

"*A Guide to Small Area Estimation*" published by the Australian Bureau of Statistics[17] has been mainly used here.

---

[12] http://www.bing.com/search?ei=UTF-8&pc=AV01&q=Small+area+estimation+in+Statistics+Canada&FROM=AVASDF&first=71&FORM=PORE.
http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3604#a3

[13] www.census.gov/hhes/www/saipe/documentation.html.
http://www.census.gov/did/www/saipe/methods/10change.html.

[14] U.K. ONS: http://www.ons.gov.uk/ons/guide-method/method-quality/survey-methodology-bulletin/.

[15] http://www.nss.gov.au/nss/home.NSF/pages/Small+Areas+Estimates?OpenDocu.

[16] http://www.bing.com/search?ei=UTF-8&pc=AV01&q=Evaluations+and+improvements+in+small+area+estimation+methodologies++Adam+Whitworth+%28edt%29%2C+University+of+Sheffield&FROM=AVASDF.

[17] http://www.nss.gov.au/nss/home.NSF/pages/Small+Areas+Estimates?OpenDocu.

5.3. Assessing the quality of small area estimates

Small area estimates are usually obtained by fitting statistical models to survey data and then applying these models to auxiliary information available for the small area population of interest. Often a number of potential or candidate models are considered involving various combinations of the auxiliary variables.

The most reliable of these candidate models is then chosen as the final model, on the basis of:

- plausibility of the model in light of previous studies or accepted wisdom;
- how well the model fits the observed data; and,
- accuracy of the small area estimates predicted from the model.

In light of this, there is a need to examine various quality diagnostics to determine which of the candidate models to use. Having chosen a model, it is then necessary to provide users with an assessment of its quality as well as the quality of the small area estimates produced from it. In doing so, ranges of diagnostics are used to assess the accuracy, validity and consistency of the small area estimates.

These include:

- a bias test that compares the small area predictions with direct estimates;
- testing whether model assumptions are met and that the model is a good fit;
- checking that small area estimates add to published state or national estimates;
- local knowledge and expert advice on the spread of estimates across small areas; and,
- relative root mean squared errors (RMSE) - in modelling these are analogous to sampling errors calculated for survey estimates.

Although these diagnostics are crucial in terms of assessing the relative performance of competing small area models, they have to be supported by good judgement from practitioners and expert advice from users.

5.4. Communication with users on quality of accepted results

From current practice we may draw conclusions that there are problems with users' communication regarding quality of accepted results. There are several propositions to improve this practice, but it is suggested to consider the following Trewin's proposition.

Trewin (1999) encouraged NSIs to make greater use of small area estimation methods to generate statistical output. However, in doing so, he emphasised that:

a) *"the estimates need to be branded differently from other official statistics (the methods and the assumptions should be described in any releases);*
b) *their validity needs to be assessed to provide user confidence;*

c) the underlying models need to be described in terms that users can understand and the validity of the underlying assumptions should be discussed with the key users;

d) their quality should be described in quantitative terms as far as possible; and

e) there should be peer review of the models by an expert as the models are very complex and the choice of methods is considerable."

The author would like to add in this section the Eurostat publication (Eurostat, 2007) devoted to data quality assessment, presenting different methods and tools.

6. Concluding remarks

Small area estimation methods have been developing significantly over the last 30 years and used partially in official statistics. Before small area estimates can be considered fully credible, carefully conducted evaluation studies are needed to check on the adequacy of the model being used. Sometimes model-dependent small area estimators turn out to be of superior quality to sample-based estimators, and this may make them seem attractive.

It seems reasonably to give some recommendations and suggestions compiled from different papers, conferences and projects related to SAE methods:

1. Good auxiliary information related to the variables of interest plays a vital role in model-based estimation. Expanded access to auxiliary data, such as census and administrative data, through coordination and cooperation among federal agencies is needed.
2. Preventive measures at the design stage may reduce the need for indirect estimators significantly.
3. Model selection and checking plays an important role. External evaluations are also desirable whenever possible.
4. Area-level models have wider scope because area-level data are more readily available. But assumption of known sampling variance is restrictive.
5. HB approach is powerful and can handle complex modelling, but caution should be exercised in the choice of priors on model parameters. Practical issues in implementing HB paradigm should be addressed.
6. Model-based estimates of area totals and means are not suitable if the objective is to identify areas with extreme population values or to identify areas that fall below or above some pre-specified level.
7. Suitable benchmarking is desirable.
8. Model-based estimates should be distinguished clearly from direct estimates. Errors in small area estimates may be more transparent to users than errors in large area estimates.

9. Proper criterion for assessing quality of model-based estimates is whether they are sufficiently accurate for the intended uses. Even if they are better than direct estimates, they may not be sufficiently accurate to be acceptable.
10. Overall program should be developed that covers issues related to sample design and data development, organization and dissemination, in addition to those pertaining to methods of estimation for small areas.

Acknowledgements

The author would like to thank his colleagues from LinkedIn: J. Bethlehem (Netherlands), J. Cochran (USA), A. Fuller (USA), D. Hedlin (Sweden), D. Kasprzyk (USA), S. Laaksonen (Finland), L. Lyberg (Sweden), D. Marker (WESTAT), Siu-Ming Tam (Australia), A. G. Turner (USA), D. Trewin (Australia), who have helped him with preparing the session "*Small Area Estimation in Official Statistics and Statistical Thinking*" held at the International Conference on *Small Area Estimation 2014*, in Poznan, Poland, 3-5 September 2014.

The author is also grateful to the anonymous referee whose comments have helped to improve the quality of this paper.

REFERENCES

AUSTRALIAN BUREAU OF STATISTICS, (2006). A Guide to Small Area Estimation – Version 1.1. Internal ABS document. available online at: http://www.nss.gov.au/nss/home.NSF/pages/Small+Areas+Estimates?OpenDocu.

BATTESE, G. E., HARTER, R. M., 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, 28–36.

BEST, N., RICHARDSON, S., THOMSON, A., (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14, 35–39.

BETHLEHEM, J. G., (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251–260.

BETHLEHEM, J. G., KERSTEN, H. M. P., (1985). On the treatment of nonresponse in sample surveys. *Journal of Official Statistics*, 1, 287–300.

BRACHA, CZ., LEDNICKI, B., WIECZORKOWSKI, R., (2003). Estimation of Data from the Polish Labour Force Surveys by poviats (counties) in 1995–2002 (in Polish), Central Statistical Office of Poland, Warsaw.

BRACKSTONE, G. J., (2002). Strategies and Approaches for Small Area Statistics, *Survey Methodology*, 28(2), 117–123.

BRACKSTONE, H., (1999). Managing Data Quality in a Statistical Agency. *Survey Methodology*, 25, 2, 129–149.

BRAKEL, J. A. VAN DEN, BETHLEHEM, J., (2008). Model-Based Estimation for Official Statistics, Statistics Netherlands, Voorburg/Heerlen.

CHAMBERS, R., BRAKEL, J. A. VAN DEN, HEDLIN, D., LEHTONEN, R., ZHANG, LI-CHUN, (2006). Future Challenges of Small Area Estimation. *Statistics in Transition*, 7, 759–769.

CHOUDRY, G. H., RAO, J. N. K., (1989). Small area estimation using models that combine time series and cross sectional data, in: Singh, A.C., Whitridge, P. (Eds.*), Proceedings of Statistics Canada Symposium on Analysis of Data in Time*, pp. 67–74.

CHOUDHRY, G. H., RAO, J. N. K., HIDIROGLOU, M. A., (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, 38, 23–29.

COCHRAN, W. G., (1977). *Sampling Techniques*, 3rd ed. New York: Wiley.

DATTA, G., (2009). Model-based approach to small area estimation. Chapter 32 in Rao C.R. and Pfeffermann D. (Eds.). Handbook of Statistics. Sample Surveys: Inference and Analysis. Vol. 29B. New York: Elsevier. (251-288).

DATTA, G. S., LAHIRI, P., MAITI, T., LU, K. L., (1999). Hierarchical Bayes estimation of unemployment rates for the U.S. states. *Journal of the American Statistical Association*, 94, 1074–1082.

DATTA, G. S., LAHIRI, P., MAITI, T., (2002). Empirical Bayes estimation of median income of four person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inferenc*e, 102, 83–97.

DEHNEL, G., (2010). *The development of micro-entrepreneurship in Poland in the light of the estimation for small areas,* Publ. University of Economics in Poznan, Poznan (in Polish).

DEHNEL, G., GOLATA, E., KLIMANEK, T., (2004). Consideration on Optimal Design for Small Area Estimation, Statistics in Transition, vol. 6, Nr 5, pp. 725–754.

DEVILLE, J., SÄRNAL, C.-E., (1992). Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, 87, 376–382.

DICK, P., (1995). Modelling net undercoverage in the 1991 Canadian census. *Survey Methodology*, 21, 45-54.

DREW, D., SINGH, M. P., CHOUDHRY, G. H., (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey, *Survey Methodology*, 8, 17–47.

ELAZAR, D., (2004), Small Area Estimation of Disability in Australia, *Statistics in Transition*, 6, 5, 667–684.

ESTEBAN, M. D., MORALES, D., PEREZ, A., SANTAMARIA, L., (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis*, 56, 2840–2855.

EURAREA, (2004). Project reference volume, deliverable D7.1.4, *Technical report, EURAREA consortium.*

EUROSTAT, (2007). *Handbook on Data Quality Assessment: Methods and Tools*, Luxembourg.

FAY, R. E., HERRIOT, R. A., (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269–277.

FULLER, W. A., (1999). Environmental surveys over time. *Journal of the Agricultural, Biological and Environmental Statistics*, 4, 331–345.

FULLER, W. A., (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5–23.

GAMBINO, J. G., Dick, P., (2000). Small Area Estimation Practice at Statistics Canada, *Statistics in Transition*, 4, 4, 597–610.

GAMBINO, J. G., SINGH, M. P., DUFOUR, J., KENNEDY, B., LINDEYER, J., (1998). *Methodology of the Canadian Labour Force Survey*, Statistics Canada.

GOLATA, E., (2004). Problems of Estimate Unemployment for Small Domains in Poland**,** *Statistics in Transition***,** 6, 5. 755–776.

GOSH, M., MEEDEN, G., (1997). Bayesian Methods for Finite Population Sampling . London: Chapman & Hall.

GHOSH, M., NATARAJAN, K., WALTER, L. A., KIM, D. H., (1999). Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping. *Journal of Statistical Planning and Inference*, 75, 305–318.

GHOSH, M., RAO, J. N. K., (1994). Small Area Estimation: An Appraisal, *Statistical Science*, 9, 55–93.

GONZALEZ-MANTEIGA, W., LOMBARDIA, M. J., MOLINA, I., MORALES, D., SANTAMARIA, L., (2010). Small area estimation under Fay-Herriot models with nonparametric estimation of heteroscedasticity. *Statistical Modelling*, 10, 2, 215–239.

HANSEN, M. H., HURWITZ, W. N., MADOW. W. G., (1953). *Sample Survey Methods and Theory*, Vol. I and II. New York: Wiley.

HEADY, P., HENNELL, S., (2001). Enhancing Small Area Estimation Techniques to Meet European Needs, *Statistics in Transition*, 5, 2, 195–203.

HEADY, P., RALPHS, M., (2004). Some Findings of the EURAREA Project – and their Implications for Statistical Policy, *Statistics in Transition*, 6, 5, 641–654.

HIDIROGLOU, M. A., (2014). Small-Area Estimation: Theory and Practice, Section on Survey Research Methods, Statistics Canada.

HIDIROGLOU, M. A., SINGH A., HAMEL M., (2007). Some Thoughts on Small Area Estimation for the Canadian Community Health Survey (CCHS). Internal Statistics Canada document.

HIDIROGLOU, M. A., SÄRNDAL, C. E., (1985). Small Domain Estimation: A Conditional Analysis, *Proceedings of the Social Statistics Section, American Statistical Association*, 147–158.

HIDIROGLOU, M. A., PATAK, Z., (2004). Domain estimation using linear regression. *Survey Methodology*, 30, 67–78.

HOLT, D., ELLIOT, D., (1991). Methods of weighting for unit non-response. *The Statistician*, 40, 333–342.

HORVITZ, D. G., THOMPSON, D. J., (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.

JIANG, J., LAHIRI, P., (2006). Mixed model prediction and small area estimation. *Test*, 15, 1–96.

KALTON, G., (2002). Models in the Practice of Survey Sampling (Revisited), *Journal of Official Statistics*, 18, 129–154.

KALTON, G., KASPRZYK, D., (1986) The treatment of missing data. *Survey Methodology*, 12, 1–16.

KALTON, G., KORDOS, J., PLATEK, R., (1993). *Small Area Statistics and Survey Designs*, Vol. I: Invited Papers; Vol. II: Contributed Papers and Panel Discussion. Central Statistical Office, Warsaw.

KISH, L., (1965). *Survey Sampling*. New York: Wiley.

KORDOS, J., (1994). Small Area Statistics in Poland (Historical Review). *Statistics in Transition*, 1, 6,783–796.

KORDOS, J., (2005). Some Aspects of Small Area Statistics and Data Quality, *Statistics in Transition*, 7, 1, 63–83.

KORDOS, J., (2006). Impact of different factors on research in small area estimation in Poland**,** „*Statistics in Transition"*, 7, 4, 863–879.

KORDOS, J., PARADYSZ, J., (2000). Some Experiments in Small Area Estimation in Poland, *Statistics in Transition*, 4, 4, 679–697.

KUBACKI, J., (2004). Application of the Hierarchical Bayes Estimation to the Polish Labour Force Survey, *Statistics in Transition*, 6, 5, 785–796.

LEHTONEN, R., SÄRNDAL, C. E., VEIJANEN, A., (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29, 33–44.

LEHTONEN, R., SÄRNDAL, C. E., VEIJANEN, A., (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, **7**, 649–673.

LEHTONEN, R., VEIJANEN, A., (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.). Handbook of Statistics. Sample Surveys: Inference and Analysis. Vol. 29B. New York: Elsevier, 219–249.

LEHTONEN, R., SÄRNDAL, C. E., VEIJANEN, A., (2009). Model calibration and generalized regression estimation for domains and small areas. Invited paper, SAE2009 Conference on Small area estimation, 29 June – 1 July, 2009, Elche, Spain.

LITTLE, R. J. A., (2004). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling, *Journal of the Royal Statistical Association*, 99, 546–556.

LITTLE, R. J. A., (2012). Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics *Journal of Official Statistics*, 28, 3, 309–334.

LONGFORD, N., (2005). *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician*, Springer.

LUNDSTRÖM, S., SÄRNDAL, C. E., (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15, 305–327.

MARHUENDA, Y. MOLINA, I., MORALES, D., (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58, 308–325.

MARKER, D. A., (2001). Producing small area estimates from national surveys: Methods for minimizing use of indirect estimators. *Survey Methodology*, 27, 183–188.

MARSHALL, A., (updated by V. Higgins, 2013). *Small area estimation using key UK surveys – An Introductory guide.* UK Data Service, University of Essex and University of Manchester.

MOLINA, I., RAO, J. N. K., (2010). Small area estimation of poverty indicators**,** *The Canadian Journal of Statistics*, 38, 3, 369–385.

MOLINA, I., SALVATI, N., PRATESI, M., (2009). Bootstrap for estimating the MSE of the Spatial EBLUP. *Computational Statistics*, 24, 441–458.

MOURA, F. A. S., MIGON, H. S., (2002). Bayesian spatial models for small area proportions. *Statistical Modelling*, 2, 3, 183–201.

NARAIN, R., (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169–174.

OPSOMER, J. D., CLAESKENS, G., RANALLI, M. G., KAUERMANN, G., BREIDT, F. J., (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society*, B70, 265–286.

PARADYSZ, J., (1998). Small Area Statistics in Poland - First Experiences and Application Possibilities, *Statistics in Transition*, 3, 5, 1003–1015.

PARADYSZ, J., DEHNEL, G., (2005). Attempts to Estimate Basic Information for Small Business in Poland, SAE2005 Conference, Jyväskylä, Finland, 28–31 August 2005.

PLATEK, R., RAO, J. N. K. SÄRNDAL, C. E., SINGH, M. P. (Eds.), (1987). *Small Area Statistics*, John Wiley & Sons, New York.

PFEFFERMANN, D., (1999). Small area estimation – big developments. Proceedings of IASS Conference on Small Area Estimation, Riga. Latvian Council of Sciences, 129–145.

PFEFFERMANN, D., (2002). Small area estimation – new developments and directions. *International Statistical Review* 70, 125–143.

PFEFFERMANN, D., (2013). New important developments in small area estimation, *Statistical Science*, 28, 1, 40–68.

PFEFFERMANN, D., BURCK, L., (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217–237.

PFEFFERMANN, D., Tiller, R., (2006). Small Area Estimation with State Space Models Subject to Benchmark Constraints. *Journal of the American Statistical Association*, 101, 1387–1897

PRASAD, N. G. N., RAO, J. N. K., (1990). "The Estimation of Mean Squared Error of Small-Area Estimators", *Journal of the American Statistical Association*, 85, 163–171.

PRASAD, N. G. N., RAO, J. N. K, (1999). On robust small area estimation using a simple random effects model. *Survey Methodology*, 25, 67–72.

PRATESI, M., SALVATI, N., (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, 17, 113–141.

RAO, J. N. K., (1999). Some recent advances in model-based small area estimation, *Survey Methodology*, 25, 2, 175–186.

RAO, J. N. K., (2003). *Small Area Estimation*, John Wiley & Sons, New Jersey.

RAO, J. N. K., (2011). Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal, *Statistical Science*, 26, 2, 240–256.

RAO, J. N. K., YU, M., (1994). Small area estimation by combining time series and cross sectional data. *Canadian Journal of Statistics*, 22, 511–528.

RIGA, (1999). *Small Area Estimation–Conference Proceedings*, Riga, Latvia, August 1999.

SÄRNDAL, C. E., SWENSSON, B., WRETMAN, J., (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SÄRNDAL, C.-E., LUNDSTRÖM, S., (2005). References, in Estimation in Surveys with Nonresponse, John Wiley & Sons.

SCHAIBLE, W. A., (1978). Choosing Weights for Composite Estimators for Small Area Statistics, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741–746.

SCHAIBLE, W. L., CASADY, R. J., (1994). The Development, Application, and Evaluation of Small Area Estimators. *Statistics in Transition*, 1, 6, 727–746.

SINGH, M. P., GAMBINO, J., MANTEL, H. J., (1994). Issues and Strategies for Small Area Data, *Survey Methodology*, 20, 3–22.

SINGH, B., SHUKLA, G., KUNDU, D., (2005). Spatio-temporal models in small area estimation. *Survey Methodology, 31*, 183–195.

SZARKOWSKI, A., WITKOWSKI, J., (1994). The Polish Labour Force Survey, *Statistics in Transition*, 1, 4, 467–483.

THOMSEN, I., HOLMOY, A. M. K., (1998). Combining data from surveys and administrative record systems. The Norwegian experience, *Inter. Satist. Rev*., 66, 201–221.

TREWIN, D., (1999). Small Area Statistics Conference, *Survey Statistician*, 41, 8–9.

TREWIN, D., (2002). The importance of a Quality Culture, *Survey Methodology*, 28, 2, 125–133.

UN, (2011). Using Administrative and Secondary Sources for Official Statistics – A Handbook of Principles and Practices, United Nations Commission for Europe.

VALLIANT, R., DORFMAN, A. H., ROYALL, R. M., (2000). *Finite Population Sampling and Inference, A Prediction Approach*. New York: Wiley.

WANG, J., FULLER, W. A., (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 463, 716–723.

WANG, J., FULLER, W. A., QU, Y., (2008). Small area estimation under a restriction. *Survey Methodology*, 34, 1, 29–36.

WOODRUFF, R. S., (1966). Use of a Regression Technique to Produce Area Breakdowns of the Monthly National Estimates of Retail Trade, *Journal of the American Statistical Association*, 61, 496–504.

WYWIAŁ, J., (2000). On precision of Horvitz–Thompson strategies, *Statistics in Transition*, 4. 5, 779–798.

YATES, F., (1981). *Sampling methods for censuses and surveys.* 4th Ed., Charles Griffin and Co., London, U.K.

YOU, Y., (2008). An integrated modelling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, 19–27.

YOU, Y., CHAPMAN, B., (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97–103.

YOU, Y., RAO, J. N. K., (2002). A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431–439.

YOU, Y., RAO, J. N. K., GAMBINO, J. G., (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach, *Survey Methodology*, 29, 25–32.

YOU, Y, DICK, P., (2004). Hierarchical Bayes Small Area Inference to the 2001 Census Undercoverage Estimation. Proceedings of the ASA Section on Government Statistics, 1836–1840.

YOU, Y., ZHOU, Q. M., (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data, *Survey Methodology*, 37, 1, 25–37.

WHITWORTH, A., (edt). Evaluations and improvements in small area estimation methodologies, University of Sheffield: http://eprints.ncrm.ac .uk/3210/1/sme_whitworth.pdf.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 133–154

A COMPARISON OF SMALL AREA AND CALIBRATION ESTIMATORS VIA SIMULATION

M. A. Hidiroglou[1], **V. M. Estevao** [2]

ABSTRACT

Domain estimates are typically obtained using calibration estimators that are direct or modified direct. They are direct if they strictly use data within the domain of interest. They are modified direct if they use both data within and outside the domain of interest. An alternative way of producing these estimates is through small area procedures. In this article, we compare the performance of these two approaches via a simulation. The population is generated using a hierarchical model that includes both area effects and unit level random errors. The population is made up of mutually exclusive domains of different sizes, ranging from a small number of units to a large number of units. We select many independent simple random samples of fixed size from the population and compute various estimates for each sample using the available auxiliary information. The estimates computed for the simulation included the Horvitz-Thompson estimator, the synthetic estimator (indirect estimate), calibration estimators, and unit level based estimators (small area estimate). The performance of these estimators is summarized based on their design- based properties.

Key words: area level, unit level, calibration estimates, small area estimates, simulation.

1. Introduction

Domain estimates at Statistics Canada are typically obtained using well-established methods based on calibration estimation. The calibration is direct or modified direct. It is direct if it is based on data within the domain of interest. It is modified direct if it is based on data within and outside the domain of interest. These methods can be viewed as design-based procedures as the variance of the resulting estimators is evaluated under the randomization distribution. The

[1] Michael A. Hidiroglou, Statistical Research and Innovation Division, Statistics Canada, 16 D, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: hidirog@yahoo.ca.
[2] Victor M. Estevao, Statistical Research and Innovation Division, Statistics Canada, 16 D, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: Victor.Estevao@statcan.gc.ca.

randomization distribution of an estimator is the distribution over all possible samples that could be selected from the target population of interest under the sampling design used to select the sample, with the population parameters considered as fixed values. Another way of producing these estimates is through small area methods. These methods are particularly important when the sample size in the domains is "small." They can improve the reliability of the direct estimates provided that the variable of interest is well correlated with auxiliary variables $\boldsymbol{x}$ that are available from administrative or other files. Small area estimation essentially combines direct estimates with model-based estimates in an optimal manner.

The model-based estimates involve known population totals (auxiliary data) and estimates of the regression between the variable of interest and the auxiliary data across the small areas. In general, these models are classified into two groups: unit level models and area level models. Unit level models are generally based on observation units (e.g., persons or companies) from the survey and auxiliary variables associated with each observation, whereas area level models are based on direct survey estimates aggregated from the unit level data and related area level auxiliary variables; see Rao (2003) for an overview of small area models. The more recent literature that covers empirical assessment of the properties of various small area and domain estimators includes Lehtonen and Veijanen (2009), Datta (2009), and Pfeffermann (2013). Lehtonen and Veijanen (2009) focused on design-based methods (calibration and regression) using auxiliary data. They reviewed the work on the extension of the linear form of the Generalized Regression Estimator (GREG) given in Särndal et al. (1992) to include logistic, multinomial logistic and mixed models for domain estimation. Datta (2009) reviewed the development of model-based procedures to obtain small area estimates. Datta focussed in particular on the theoretical properties of the resulting estimators. Pfeffermann (2013) reviewed both design-based and model-based procedures, as well as recent developments in these two procedures.

Domain estimates are currently obtained via design-based procedures at Statistics Canada. However, the increasing requirement for producing estimates for "small domains" has encouraged the need to adopt model-based procedures. A SAS-based prototype (Estevao et al. 2014) has been recently developed at Statistics Canada to respond to these requirements. The prototype currently incorporates two well-known methods initially developed by Fay and Herriot (1979) for area level estimation, and Battese, Harter, and Fuller (1988) for unit level estimation. Although the theoretical properties of the estimators included in the prototype are known, they were investigated via a simulation. In the simulations, we looked at the properties of estimators of domain totals. We compared model-based small area estimators with traditional estimators through simulation. The latter included the Horvitz-Thompson estimator, two calibration estimators, the modified regression estimator and the synthetic estimator. The small area estimators are the EBLUP and Pseudo EBLUP estimators based on a unit level model. More details on all of these estimators are given in section 2.

The simulation setup and results are reported in section 3. Section 4 provides a few conclusions from our findings.

2. Sample design

Large scale surveys are designed to satisfy reliability requirements for some subsets (domains) of the population. Examples of these subsets include partitions below the level of the initial geographical / industrial detail requested by the client. If such subsets are required before the sample is selected, then such domains are labelled as *planned domains* (Singh, Gambino and Mantel 1994). Such planned domains will have some of the sample allocated to them to obtain unbiased estimates with the required precision using direct estimation procedures. If these domains are identified after the sample has been selected, they will be known as *unplanned domains*. Note that, in any event, unplanned domains will exist for most surveys. An example taken from business surveys is a change of industry during data collection. A business initially classified as industry A becomes industry B. Such a business would be tabulated as part of the businesses of type B, but would retain its original sampling weight. Another example, taken from household surveys, would be the arbitrary production of estimates below a geographical level that was not part of the allocation process of the sample. Traditional or small area estimators can be used for either planned or unplanned domains.

As domain estimation for most surveys at Statistics Canada is mostly of the unplanned type, we have designed our simulation to reflect this tendency: that is no units are allocated to them prior to sample selection. Domain estimates are produced after sample selection, and the number of sampled units falling in each domain is a random variable. Our simulation reflects this point, and we used the simplest sample design to carry it out. We drew repeated samples *s* of size *n* from the population *U* of size *N* using simple random sampling without replacement. The weight associated with unit $j \in U$ is denoted as w_j. Let s_d, $d = 1,2,...,D$, be the portion of the sample *s* that overlaps with domain U_d (of known size N_d). Let the realized sample size in domain U_d be n_d. The survey design weight associated with a unit $j \in U_d$ is w_j. The data in the population are denoted as $\left(y_j, \boldsymbol{x}_j\right)$ for each element $j \in U$. The *y* variable is the one of interest, while $\boldsymbol{x}$ is the vector of auxiliary data. Computation of domain statistics can be obtained using the operators (i.e.: mean and variance) in regular estimation via the following transformation. In domain U_d, we denote the variable of interest as y_{dj} where $y_{dj} = y_j$ if $j \in U_d$ and 0 otherwise. The associated vector of auxiliary variables is defined as $\boldsymbol{x}_{dj}$ where $\boldsymbol{x}_{dj} = \boldsymbol{x}_j$ if $j \in U_d$ and $\mathbf{0}$ otherwise.

The objective of the present study is to compare the properties of model-based small area estimators for domains with those traditionally used in survey estimation. We considered seven estimators of the domain total $Y_d = \sum_{j \in U} y_{dj}$: four are traditional estimators and three are small area estimators. We first present the traditional estimators.

2.1. Traditional estimators

Horvitz-Thompson: The Horvitz-Thompson estimator $\hat{Y}_{d\,HT}$, $d = 1, 2, ..., D$, uses no auxiliary information. It is defined as $\hat{Y}_{d\,HT} = \sum_{j \in s} w_j y_{dj}$ if $n_d > 0$ and 0 otherwise. We set $\hat{Y}_{d\,HT}$ to 0 if there are no sampled units in the domain, ensuring unbiased estimation over all samples *s* drawn from *U*. Although this estimator is unbiased, it produces inefficient estimates.

Calibration Estimators: We consider two calibration estimators, $\hat{Y}_{d\,CALU_d}$ and $\hat{Y}_{d\,CALU}$, that use auxiliary information at different levels. They are applications of calibration given in Deville and Särndal (1992) adapted to domain estimation. The direct estimator $\hat{Y}_{d\,CALU_d}$ uses auxiliary information at the domain level, while the modified direct estimator $\hat{Y}_{d\,CALU}$ uses information at the population level. Estimator $\hat{Y}_{d\,CALU_d}$ is known to be more efficient than $\hat{Y}_{d\,CALU}$. However, estimator $\hat{Y}_{d\,CALU_d}$ has some drawbacks. It is not always possible to obtain auxiliary information at the domain level. Even if this information is available, we cannot produce estimates using $\hat{Y}_{d\,CALU_d}$ if there are no sample units in the domain. Furthermore, this estimator can produce erratic values when there are only a few units in the domain. To prevent this, we need to make sure that the number of units in the domain is larger than the number of auxiliary variables. As a minimal requirement, given that there are two auxiliary variables (intercept, *x*), $\hat{Y}_{d\,CALU_d}$ can be estimated only if there are 3 or more units in a domain. Otherwise, we cannot produce a value, and we set it to missing. This means that we only work with a subset of all possible samples. If we set the value of $\hat{Y}_{d\,CALU_d}$ to 0 when there is an insufficient number of observations n_d in domain U_d, this would result in a biased estimator. As for $\hat{Y}_{d\,CALU}$, when there are no sample units in the domain, we set the value of this estimator to 0. This ensures that it is

approximately design unbiased for the domain total. Estimator $\hat{Y}_{d\,CALU_d}$, $d = 1, 2, ..., D$, is given by:

$$\hat{Y}_{d\,CALU_d} = \begin{cases} \sum_{j \in s} w_j y_{dj} + (\boldsymbol{X}_d - \hat{\boldsymbol{X}}_{d\,HT})^T \hat{\boldsymbol{\beta}}_{d\,CALU_d} & \text{if } n_d \geq 3 \\ . \text{ (missing)} & \text{if } n_d < 3 \end{cases}$$

with $\boldsymbol{X}_d = \sum_{j \in U} \boldsymbol{x}_{dj}$, $\hat{\boldsymbol{X}}_{d\,HT} = \sum_{j \in s} w_j \boldsymbol{x}_{dj}$ and

$$\hat{\boldsymbol{\beta}}_{d\,CALU_d} = \left(\sum_{j \in s} \frac{w_j \boldsymbol{x}_{dj} \boldsymbol{x}_{dj}^T}{c_j} \right)^{-1} \sum_{j \in s} \frac{w_j \boldsymbol{x}_{dj} y_{dj}}{c_j}.$$

Estimator $\hat{Y}_{d\,CALU}$, $d = 1, 2, ..., D$, is given by:

$$\hat{Y}_{d\,CALU} = \begin{cases} \sum_{j \in s} w_j y_{dj} + (\boldsymbol{X} - \hat{\boldsymbol{X}}_{HT})^T \hat{\boldsymbol{\beta}}_{CALU} & \text{if } n_d > 0 \\ 0 & \text{if } n_d = 0 \end{cases}$$

with $\boldsymbol{X} = \sum_{j \in U} \boldsymbol{x}_j$, $\hat{\boldsymbol{X}}_{HT} = \sum_{j \in s} w_j \boldsymbol{x}_j$ and $\hat{\boldsymbol{\beta}}_{CALU} = \left(\sum_{j \in s} \frac{w_j \boldsymbol{x}_j \boldsymbol{x}_j^T}{c_j} \right)^{-1} \sum_{j \in s} \frac{w_j \boldsymbol{x}_j y_j}{c_j}$.

Modified Regression (REG): The modified regression estimator $\hat{Y}_{d\,REG}$, $d = 1, 2, ..., D$, is due to Woodruff (1966). It is of interest as it was used to produce area breakdowns of the monthly national estimates of US Census Bureau retail trade survey. Note that it is a modified direct estimator. Singh and Mian (2003) points out that it can be viewed as a calibration estimator $\sum_s \tilde{w}_{dj} y_j$, where the calibration weight $\tilde{w}_{dj}$ is obtained by minimizing the chi-squared distance $\sum_s c_j \left(w_j a_{dj} - \tilde{w}_{dj} \right) / w_j$, subject to the constraints $\sum_s \tilde{w}_{dj} \boldsymbol{x}_j = \boldsymbol{X}_d$: here a_{dj} is the domain indicator variable. Estimator $\hat{Y}_{d\,REG}$ is design-unbiased as the overall sample size increases. It is given by:

$$\hat{Y}_{d\,REG} = \begin{cases} \sum_{j \in s_d} w_{dj} y_{dj} + (\boldsymbol{X}_d - \hat{\boldsymbol{X}}_{d\,HT})^T \hat{\boldsymbol{\beta}}_{REG} & \text{if } n_d > 0 \\ \boldsymbol{X}_d^T \hat{\boldsymbol{\beta}}_{REG} & \text{if } n_d = 0 \end{cases}$$

with $\boldsymbol{X}_d = \sum_{j \in U_d} \boldsymbol{x}_{dj}$, $\hat{\boldsymbol{X}}_{d\,HT} = \sum_{j \in s} w_j \boldsymbol{x}_{dj}$ and $\hat{\boldsymbol{\beta}}_{REG} = \left(\sum_{j \in s} \frac{w_j \boldsymbol{x}_j \boldsymbol{x}_j^T}{c_j} \right)^{-1} \sum_{j \in s} \frac{w_j \boldsymbol{x}_j y_j}{c_j}$.

The c_j term, $c_j > 0$, associated with the estimators that use auxiliary data reflects that the error terms e_j in the implied working model are distributed independently with mean zero and variance $c_j^2 \sigma_e^2$.

2.2. Small area estimators

The simplest small area estimator is the synthetic estimator (SYN), $\hat{Y}_{d\,SYN}$, $d = 1, 2, ..., D$. It is given by $\hat{Y}_{d\,SYN} = \boldsymbol{X}_d^T \hat{\boldsymbol{\beta}}_{SYN}$ where $\boldsymbol{X}_d = \sum_{j \in U_d} \boldsymbol{x}_{dj}$ and $\hat{\boldsymbol{\beta}}_{SYN} = \left(\sum_{j \in s} \frac{w_j \boldsymbol{x}_j \boldsymbol{x}_j^T}{c_j} \right)^{-1} \sum_{j \in s} \frac{w_j \boldsymbol{x}_j y_j}{c_j}$. This estimator is design-biased, given by $Bias\,(\hat{Y}_{d\,SYN}) \approx \boldsymbol{X}_d^T \boldsymbol{B} - Y_d$, where $\boldsymbol{B} = \left(\sum_{j \in U} \frac{\boldsymbol{x}_j \boldsymbol{x}_j^T}{c_j} \right)^{-1} \sum_{j \in U} \frac{\boldsymbol{x}_j y_j}{c_j}$ is the population regression vector.

The next two small area estimators are based on a hierarchical model given by:

$$y_{dj} = \boldsymbol{x}_{dj}^T \boldsymbol{\beta} + v_d + e_{dj}, \tag{1}$$

where $v_d \overset{iid}{\sim} N(0, \sigma_v^2)$, $e_{dj} \overset{iid}{\sim} N(0, c_{dj}^2 \sigma_e^2)$, and c_{dj} accounts for possible heterogeneity of the e_{dj} residuals.

In our application of this model, the areas are our domains of interest. The quantity $\boldsymbol{x}_{dj}^T \boldsymbol{\beta}$ is the fixed effect which is assumed to be a linear combination of the auxiliary variables $\boldsymbol{x}_{ij}$. The residuals v_d and e_{dj} are respectively the random effect for the area *d* and the random errors for unit *j* in area *d*. The term c_{dj}^2 translates to $a_{dj} = c_{dj}^{-2}$ in the various formulas that follow.

Empirical Best Linear Unbiased Predictor (EBLUP): This estimator denoted as $\hat{Y}_{d\,EBLUP}$, $d = 1,2,...,D$, is given in Rao (2003, p.136). It is an extension of the Battese, Harter, and Fuller (1988) estimator when the error structure of the residuals is not homogeneous. It is given by:

$$\hat{Y}_{d\,EBLUP} = \begin{cases} N_d \{ \bar{\boldsymbol{X}}_d^T \hat{\boldsymbol{\beta}}_{EBLUP} + \hat{\gamma}_{da} (\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T \hat{\boldsymbol{\beta}}_{EBLUP}) \} & \text{if } n_d > 0 \\ \boldsymbol{X}_d^T \hat{\boldsymbol{\beta}}_{EBLUP} & \text{if } n_d = 0 \end{cases}$$

The terms making up $\hat{Y}_{d\,EBLUP}$ include N_d, $\boldsymbol{X}_d$, $\hat{\gamma}_{da}$ $\bar{y}_{da}$ $\bar{\boldsymbol{x}}_{da}$, and $\hat{\boldsymbol{\beta}}_{EBLUP}$. These terms are defined as follows: $\bar{y}_{da} = \dfrac{\sum_{j \in s_d} a_{dj} y_{dj}}{\sum_{j \in s_d} a_{dj}}$, $\bar{\boldsymbol{x}}_{da} = \dfrac{\sum_{j \in s_d} a_{dj} \boldsymbol{x}_{dj}}{\sum_{j \in s_d} a_{dj}}$, and $\hat{\gamma}_{da} = \dfrac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \, \delta_{da}^2}$ where $\delta_{da}^2 = \dfrac{1}{\sum_{j \in s_d} a_{dj}}$. The estimated regression vector is given by:

$$\hat{\boldsymbol{\beta}}_{EBLUP} = \left(\sum_{d=1}^{D} \sum_{j \in s_d} a_{dj} (\boldsymbol{x}_{dj} - \hat{\gamma}_{da} \bar{\boldsymbol{x}}_{da}) \boldsymbol{x}_{dj}^T \right)^{-1} \sum_{d=1}^{D} \sum_{j \in s_d} a_{dj} (\boldsymbol{x}_{dj} - \hat{\gamma}_{da} \bar{\boldsymbol{x}}_{da}) y_{dj} \, .$$

This estimator is not design consistent, unless the sampling design is self-weighting.

Pseudo-EBLUP (PEBLUP): This estimator denoted as $\hat{Y}_{d\,PEBLUP}$, $d = 1,2,...,D$, is an extension of the Pseudo-EBLUP estimator given in You and Rao (2002). It accounts for the heterogeneity of the e_{dj} residuals in model (1). It includes the survey weights w_j, $j \in s$, in the regression coefficient and the parameter estimate.

$$\hat{Y}_{d\,PEBLUP} = \begin{cases} N_d \{ \bar{\boldsymbol{X}}_d^T \hat{\boldsymbol{\beta}}_{PEBLUP} + \hat{\gamma}_{dw} (\bar{y}_{dw} - \bar{\boldsymbol{x}}_{dw}^T \bar{y}_{dw}) \} & \text{if } n_d > 0 \\ \boldsymbol{X}_d^T \hat{\boldsymbol{\beta}}_{PEBLUP} & \text{if } n_d = 0 \end{cases}$$

The terms making up $\hat{Y}_{d\,PEBLUP}$ include N_d, $\boldsymbol{X}_d$, $\bar{y}_{dw}$, $\bar{\boldsymbol{x}}_{dw}$, $\bar{y}_{dw}$, and $\hat{\boldsymbol{\beta}}_{PEBLUP}$.

These terms are defined as follows: $\bar{y}_{dw} = \dfrac{\sum_{j \in s_d} w_j y_{dj}}{\sum_{j \in s_d} w_j}$, $\bar{\boldsymbol{x}}_{dw} = \dfrac{\sum_{j \in s_d} w_j \boldsymbol{x}_{dj}}{\sum_{j \in s_d} w_j}$,

$\hat{\gamma}_{dw} = \dfrac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \, \delta_{dw}^2}$, where $\delta_{dw}^2 = \dfrac{\sum_{j \in s_d} w_j^2 \, / \, a_{dj}}{\left(\sum_{j \in s_d} w_j \right)^2}$ for $d = 1,2,...,D$.

The estimated regression vector is given by:

$$\hat{\boldsymbol{\beta}}_{PEBLUP} = \left(\sum_{d=1}^{D} \sum_{j \in s_d} w_j a_{dj} (\boldsymbol{x}_{dj} - \hat{\gamma}_{dwa} \bar{\boldsymbol{x}}_{dwa}) \boldsymbol{x}_{ij}^T \right)^{-1} \sum_{d=1}^{D} \sum_{j \in s_d} w_j a_{dj} (\boldsymbol{x}_{dj} - \hat{\gamma}_{dwa} \bar{\boldsymbol{x}}_{dwa}) y_{dj}$$

with $\bar{\boldsymbol{x}}_{dwa} = \dfrac{\sum_{j \in s_d} w_j a_{dj} \boldsymbol{x}_{dj}}{\sum_{j \in s_d} w_j a_{dj}}$ and $\hat{\gamma}_{dwa} = \dfrac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \delta_{dwa}^2}$

where $\delta_{dwa}^2 = \dfrac{\sum_{j \in s_d} \dfrac{(w_j a_{dj})^2}{a_{dj}}}{\left(\sum_{j \in s_d} w_j a_{dj} \right)^2}$.

This estimator is design consistent.

3. Simulation

Surveys produced at Statistics Canada can be as simple as stratified one stage simple random sampling designs typically used for business surveys to the more complex stratified multi-stage design with unequal selection probabilities at each stage typically used for household surveys. We opted for a single stage simple random sample selected from the population, as it is a simplification of the sample designs used for business surveys. Had we chosen a sampling design with unequal weights, we would have had to account for the possible impact of informative sampling on the small area estimators using the procedure given in Pfeffermann and Sverchkov (2007). Verret, Rao, and Hidiroglou (2015) used a simpler procedure than the one given in Pfeffermann and Sverchkov (2007). Their procedure accounted for unequal selection probabilities for model-based small area estimators by incorporating them into the model. Their simulation used a design-model (*pm*) approach. Their results showed that incorporating the unequal selection probabilities significantly improved the performance (average absolute bias and average RMSE) of EBLUP, but had marginal impact on PEBLUP.

3.1 Population Generation and Sample Selection

A population *U* consisting of 4,640 units was created by generating data (x_{ij}, y_{ij}) for three separate subsets of the population (groups) with different intercepts and slopes. Each group was split into mutually exclusive and exhaustive domains as follows: Group 1 was split into nine domains $U_1,...,U_9$; Group 2 was split into ten domains $U_{10},...,U_{19}$; and Group 3 was split into ten domains $U_{20},...,U_{29}$. The three groups resulted in a total of *D*=29 domains that were mutually exclusive and exhaustive. The number of units in each domain, N_d, was

allocated in a monotonic manner: domain U_1 had 20 units; domain U_2 had 30 units; and domain U_{29} had 300 units. In our simulation the auxiliary data $\boldsymbol{x}$ consisted of two auxiliary variables. The first one had the fixed value of one to represent the intercept in the model. The second one, x, represented the available auxiliary data in the population. The auxiliary variable x in each group was generated from a $Gamma\,(\alpha = 5, \beta = 10)$ distribution with mean $\alpha\beta = 50$ and variance $\alpha\beta^2 = 500$. The variable of interest y was generated using the model

$$y_{dj} = \beta_{0,\ell} + \beta_{1,\ell}\, x_{dj} + v_d + e_{dj} : \ \ell = 1,2,3 \tag{2}$$

where $v_d \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{dj} \overset{iid}{\sim} N(0, c_{dj}^2\, \sigma_e^2)$.

We used $\sigma_v^2 = \sigma_e^2 = 20^2 = 400$ and set c_{dj}^2 equal to x_{dj}. The following table summarizes how the population was split into the three groups of domains.

Table 1. Groups, associated domains and regression parameters

Group (ℓ)	***Domains in Group***	$\beta_{0,\ell}$	$\beta_{1,\ell}$
1	U_d for d= 1,..., 9	200	30
2	U_d for d=10,...,19	300	20
3	U_d for d =20,...,29	400	10

A plot of the generated population is shown in Figure 1. The units in the groups are shown respectively in green, blue and yellow. The three regression lines are shown in red. Without the colours to identify the groups, one might be inclined to think that the population was generated under a model with a single auxiliary variable (one intercept and slope) as shown in the inset.

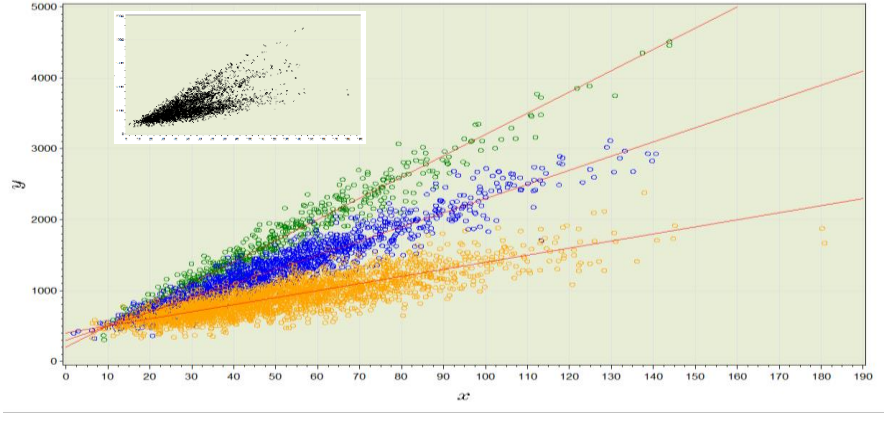


Figure 1. Plot of y vs. x for population in the simulation study

We ran two separate simulation "runs" to reflect that two possible models could be fitted for the selected samples. We denote these simulations as runs 1 and 2. In the first run (simulation run 1), we assumed that the model could be fitted using $\boldsymbol{x}_{dj} = (1, x_{dj})$ as auxiliary data; this is not correct as the population was generated on the basis of three different regressions. In the second run (simulation run 2), we acknowledged that there were three separate models and used a set of auxiliary variables reflecting the manner in which the population values were generated; this fit is correct. This meant using a set of dummy-coded auxiliary variables defined as follows for each unit:

$$\boldsymbol{x}_{dj}^{T} = \begin{cases} \left(1,0,0,x_{dj},0,0\right) & \text{if } j \in U_d \in \text{group 1} \\ \left(0,1,0,0,x_{dj},0\right) & \text{if } j \in U_d \in \text{group 2} \\ \left(0,0,1,0,0,x_{dj}\right) & \text{if } j \in U_d \in \text{group 3} \end{cases} \tag{3}$$

In the small area estimation model given by equation (1), the use of this $\boldsymbol{x}_{dj}$ implies the following regression coefficient $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6\right)^T$ for the fixed effects. For the synthetic estimator and the calibration estimators, we set $c_{ij} = x_{ij}$ to reflect the heterogeneity of the model errors.

Each simulation run involved the selection of R=100,000 independent samples and the computation of various estimates for each sample. Each sample was a simple random sample s of size n selected without replacement from U. We used sample sizes n=232 (5%), n=464 (10%), n=696 (15%) and n=928 (20%), where the sampling fractions are indicated in brackets. These are within the range of the sampling fractions typically used by business surveys.

The sample units in domain U_d are denoted by s_d with $s = \bigcup_{d=1}^{D} s_d$. We observed n_d units in U_d where $0 \le n_d \le N_d$ and $n = \sum_{d=1}^{D} n_d$. Under simple random sampling without replacement, the n_d follow a multivariate hypergeometric distribution with probability mass function $\prod_{d=1}^{D} \binom{N_d}{n_d} \Big/ \binom{N}{n}$.

The following table shows the probability of observing $n_d = 0$, $n_d = 1$ or $n_d = 2$ in the three smallest domains when the sample size n is 232.

Table 4. Probabilities in the 3 smallest domains when n = 232

Probability	U_1 with $N_1 = 20$	U_2 with $N_2 = 30$	U_3 with $N_3 = 40$
Prob $(n_d = 0)$	0.358	0.214	0.127
Prob $(n_d = 1)$	0.378	0.339	0.271
Prob $(n_d = 2)$	0.189	0.260	0.279

Using table 4, for the smallest domain U_1, $\hat{Y}_{d\,HT}$ and $\hat{Y}_{d\,CALU}$ would be equal to zero about 36% of the time. Note that this probability decreases rapidly as the domain population size N_d increases. Since we require $n_d \geq 3$, we cannot produce an estimate for $\hat{Y}_{d\,CALU_d}$ in approximately 92.5% of the samples selected in the smallest domain U_1. This probability decreases rapidly as the domain population size N_d increases.

3.2. Simulation statistics

For each selected sample in each simulation run $r = 1,...,R$ (R=100,000), we computed estimates of Y_d for the seven estimators. Denote $\hat{Y}_{d\,EST}^{(r)}$ as the estimate produced for the r^{th} sample, $r = 1, 2, ... R$, where the subscript '*EST*' is a placeholder for any one of the seven estimators. For each domain d=1,...,29, we computed the bias as:

$$Bias(\hat{Y}_{d\,EST}) = R^{-1}\sum_{r=1}^{R}\hat{Y}_{d\,EST}^{(r)} - Y_d$$

and the mean squared error as

$$MSE(\hat{Y}_{d\,EST}) = R^{-1}\left(\sum_{r=1}^{R}\hat{Y}_{d\,EST}^{(r)} - Y_d\right)^2.$$

For each estimator, $\hat{Y}_{d\,EST}$, we also computed the following summary statistics across all domains and simulated samples. These were the average absolute relative bias, the average coefficient of variation and the average relative efficiency denoted as $\overline{ARB}(\hat{Y}_{EST})$, $\overline{CV}(\hat{Y}_{EST})$ and $\overline{RE}(\hat{Y}_{EST})$ respectively.

These were computed as follows:

$$\overline{ARB}(\hat{Y}_{EST}) = \frac{1}{D}\sum_{d=1}^{D} ARB(\hat{Y}_{d\,EST}) \quad \text{where} \quad ARB(\hat{Y}_{d\,EST}) = \left| \frac{Bias(\hat{Y}_{d\,EST})}{Y_d} \right|$$

$$\overline{CV}(\hat{Y}_{EST}) = \frac{1}{D}\sum_{d=1}^{D} CV(\hat{Y}_{d\,EST}) \quad \text{where} \quad CV(\hat{Y}_{d\,EST}) = \frac{\sqrt{MSE(\hat{Y}_{d\,EST})}}{Y_d}$$

$$\overline{RE}(\hat{Y}_{EST}) = \sqrt{\frac{\overline{MSE}(\hat{Y}_{HT})}{\overline{MSE}(\hat{Y}_{EST})}} \quad \text{where} \quad \overline{MSE}(\hat{Y}_{EST}) = \frac{1}{D}\sum_{d=1}^{D} MSE(\hat{Y}_{d\,EST}) \tag{4}$$

The statistic $\overline{RE}(\hat{Y}_{EST})$ measures the average efficiency of each estimator relative to the Horvitz-Thompson estimator. Since $\hat{Y}_{d\,HT}$ is known to have the least efficiency among these seven estimators, this measure is a number larger than or equal to 1.

3.3. Simulation Results

Tables 5, 6 and 7 show the differences between the two runs using the summary statistics described in the previous section. The results are discussed after each of these three tables for runs 1 and 2.

Table 5. Average Absolute Relative Bias $\overline{ARB}(\hat{Y}_{dEST})$

		Traditional Domain Estimators				**Small Area Estimators**		
Sample Size	**Run**	$\hat{Y}_{dHT}$	$\hat{Y}_{d\,CALU_d}$	$\hat{Y}_{d\,CALU}$	$\hat{Y}_{d\,REG}$	$\hat{Y}_{dSYN}$	$\hat{Y}_{dEBLUP}$	$\hat{Y}_{dPEBLUP}$
232	1	0.12	0.16	0.15	0.19	24.18	7.58	4.12
	2	0.11	0.15	0.36	0.05	1.33	1.07	1.08
464	1	0.08	0.08	0.09	0.10	24.18	6.71	2.24
	2	0.06	0.07	0.19	0.02	1.33	0.95	0.96
696	1	0.06	0.05	0.06	0.06	24.18	6.43	1.52
	2	0.05	0.04	0.11	0.02	1.33	0.84	0.86
928	1	0.06	0.03	0.06	0.04	24.18	6.29	1.14
	2	0.05	0.03	0.09	0.01	1.33	0.76	0.77

Model does not fit (**Run 1**): The small area estimators $\hat{Y}_{dSYN}$, $\hat{Y}_{dEBLUP}$, and $\hat{Y}_{dPEBLUP}$ have the largest $\overline{ARB}$s. In particular, $\hat{Y}_{dSYN}$ has the highest $\overline{ARB}$. The $\overline{ARB}$ decreases as the sample size increases for $\hat{Y}_{dEBLUP}$ and $\hat{Y}_{dPEBLUP}$, whereas it remains constant (as expected) for $\hat{Y}_{dSYN}$. The $\overline{ARB}$ associated with $\hat{Y}_{dPEBLUP}$ decreases more rapidly than the one associated with $\hat{Y}_{dEBLUP}$ as the sample size increases. The $\overline{ARB}$ associated with the traditional domain estimators is quite small: it also decreases as the sample size increases.

Model fits (**Run 2**): The $\overline{ARB}$s associated with the small area estimators have significantly decreased. However, they are still higher than those associated with the traditional domain estimators. $\hat{Y}_{d\,REG}$ has the smallest $\overline{ARB}$ amongst all the estimators. As noted in run 1, the $\overline{ARB}$ decreases as the sample size increases for all the estimators.

Table 6. Average Coefficient of Variation $\overline{CV}(\hat{Y}_{dEST})$

		Traditional Domain Estimators			**Small Area Estimators**			
Sample Size	**Run**	$\hat{Y}_{dHT}$	$\hat{Y}_{d\,CALU_d}$	$\hat{Y}_{d\,CALU}$	$\hat{Y}_{d\,REG}$	$\hat{Y}_{dSYN}$	$\hat{Y}_{dEBLUP}$	$\hat{Y}_{dPEBLUP}$
232	1	42.79	6.57	42.81	12.82	24.27	9.90	7.93
	2	42.77	6.39	42.04	4.47	2.17	2.22	2.21
464	1	29.41	4.09	29.40	8.84	24.22	8.18	5.36
	2	29.45	4.36	28.64	3.10	1.82	1.77	1.77
696	1	23.33	2.98	23.32	7.02	24.21	7.49	4.20
	2	23.36	3.01	22.64	2.46	1.69	1.54	1.55
928	1	19.61	2.38	19.59	5.90	24.20	7.10	3.49
	2	19.61	2.35	18.96	2.07	1.61	1.39	1.40

Model does not fit (**Run 1**): Estimators $\hat{Y}_{dHT}$ and $\hat{Y}_{d\,CALU}$ have the highest $\overline{CV}$ among all estimators; their $\overline{CV}$s are quite comparable, implying that auxiliary

data used at the population level in $\hat{Y}_{d\,CALU}$ has no impact on improving the reliability of the estimator at the domain level. The synthetic estimator $\hat{Y}_{dSYN}$ also has a high $\overline{CV}$ that remains constant no matter what the sample size is. The calibration estimator $\hat{Y}_{d\,CALU_d}$ has the lowest $\overline{CV}$ for all sample sizes. The ranking from low to high of the remaining three estimators is $\hat{Y}_{d\,PEBLUP}$, $\hat{Y}_{dEBLUP}$ and $\hat{Y}_{d\,REG}$. Note that the $\overline{CV}$ of $\hat{Y}_{d\,REG}$ decreases quite rapidly as compared to the other estimators. The reliability of all the estimators improves as the sample size increases.

Model fits (**Run 2**): The $\overline{CV}$s are smaller than those obtained in run 1 for all estimators except for $\hat{Y}_{dHT}$ and $\hat{Y}_{d\,CALU}$. This is expected as both estimators do not profit from the auxiliary data. These two estimators are still the ones with the highest $\overline{CV}$s. As expected, because the model fits well, all three small area estimators $\hat{Y}_{dSYN}$, $\hat{Y}_{dEBLUP}$ and $\hat{Y}_{dPEBLUP}$ have reasonable $\overline{CV}$s. The modified regression estimator $\hat{Y}_{d\,REG}$ performs better than the calibration at the domain level $\hat{Y}_{d\,CALU_d}$: the reverse was true when the model was incorrect (run 1).

Table 7. Average Relative Efficiency $\overline{RE}(\hat{Y}_{dEST})$

		Traditional Domain Estimators			**Small Area Estimators**			
Sample Size	**Run**	$\hat{Y}_{dHT}$	$\hat{Y}_{d\,CALU_d}$	$\hat{Y}_{d\,CALU}$	$\hat{Y}_{d\,REG}$	$\hat{Y}_{dSYN}$	$\hat{Y}_{dEBLUP}$	$\hat{Y}_{dPEBLUP}$
232	1	1.00	6.43	1.00	3.48	1.50	4.04	5.48
	2	1.00	6.57	1.03	8.48	13.23	13.97	13.96
464	1	1.00	7.39	1.00	3.47	1.03	3.25	5.59
	2	1.00	7.18	1.04	8.43	9.84	11.72	11.64
696	1	1.00	7.87	1.00	3.47	0.82	2.75	5.62
	2	1.00	7.85	1.04	8.41	8.03	10.67	10.56
928	1	1.00	8.07	1.00	3.46	0.69	2.40	5.64
	2	1.00	8.10	1.04	8.40	6.84	10.06	9.95

Note: The higher the number the more efficient the estimator relative to the HT estimator. Recall that run 1 represents the results when the model does not fit, whereas run 2 represents the results when the model fits.

Model does not fit (**Run 1**): The ranking of the estimators (from highest $\overline{RE}$ to lowest $\overline{RE}$) is as follows: $\hat{Y}_{d\,CALU_d}$, $\hat{Y}_{dPEBLUP}$, $\hat{Y}_{dEBLUP}$, $\hat{Y}_{d\,REG}$, $\hat{Y}_{dSYN}$, and $\hat{Y}_{d\,CALU}$. The traditional domain estimator $\hat{Y}_{d\,CALU_d}$ is doing the best, but it is closely followed by the two small area estimators $\hat{Y}_{dPEBLUP}$ and $\hat{Y}_{dEBLUP}$. As the sample size increases, there is a dichotomy in terms of $\overline{RE}$. The relative efficiency increases for $\hat{Y}_{d\,CALU_d}$ and $\hat{Y}_{dPEBLUP}$, whereas it decreases for $\hat{Y}_{d\,REG}$, $\hat{Y}_{dSYN}$, and $\hat{Y}_{dEBLUP}$. There is no change to $\hat{Y}_{d\,CALU}$ as the auxiliary information is not useful at the domain level.

Model fits (**Run 2**): The ranking of the estimators (from highest $\overline{RE}$ to lowest $\overline{RE}$) has changed with respect to run 1. It is now $\hat{Y}_{dEBLUP}$, $\hat{Y}_{dPEBLUP}$, $\hat{Y}_{dSYN}$, $\hat{Y}_{d\,REG}$, $\hat{Y}_{d\,CALU_d}$, and $\hat{Y}_{d\,CALU}$. The small area estimators are clearly more efficient than the traditional estimators. The relative efficiency increased for all estimators - maximum is now 14 versus 8 obtained in run 1. Once more, as the sample size increases, there is a dichotomy in terms of $\overline{RE}$.

Another way to summarize the behaviour of the various estimators is graphically. We summarized the average absolute relative bias, $ARB(\hat{Y}_{d\,EST})$, and the average coefficient of variation, $CV(\hat{Y}_{d\,EST})$, within each domain $d = 1, 2, \ldots, D$, where $D = 29$.

Figures 2a and 2b display two typical graphs of the absolute relative bias over the domains for the two simulation runs. These graphs show the results for the sample size of 464. Similar results were obtained for the other sample sizes. We can see that the absolute relative bias of $\hat{Y}_{d\,SYN}$, $\hat{Y}_{d\,EBLUP}$ and $\hat{Y}_{d\,PEBLUP}$ is greatly reduced when we specify the 'correct' auxiliary variables in the underlying model. In the first run, the small area estimators show a 'drop' and a 'rise' between the groups of domains. This can be explained. The overall model fitted using $\boldsymbol{x}_{ij} = (1, x_{ij})$ produces a regression which is close to the underlying model for the second group of domains. Therefore, the differences are small for the second group of domains. However, this overall model is quite different from the one used to generate the population in the first and third groups of domains.

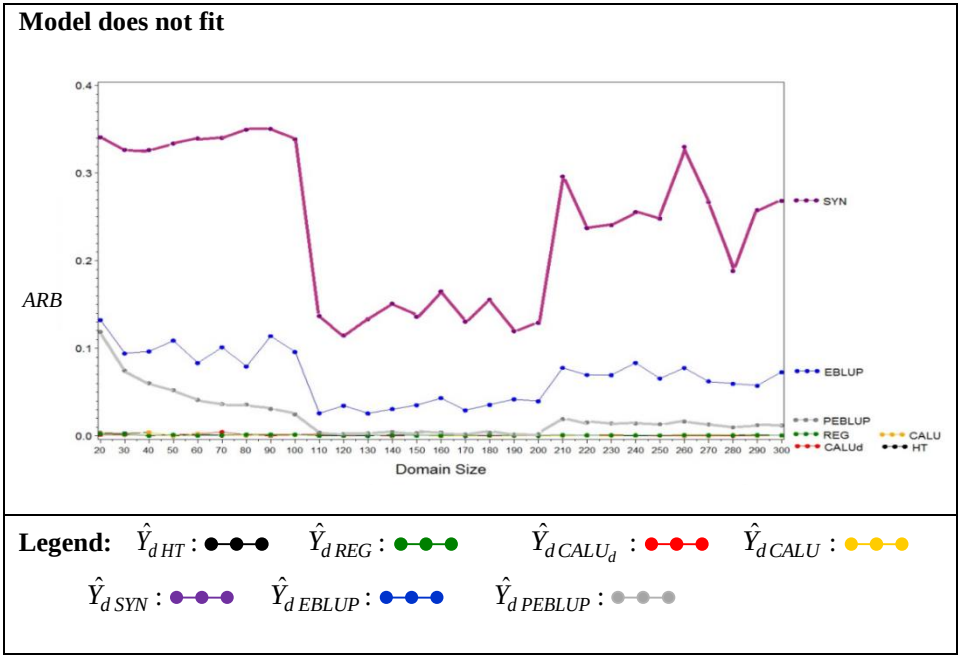


Figure 2a. Plots of the absolute relative bias of the estimators for sample size 464

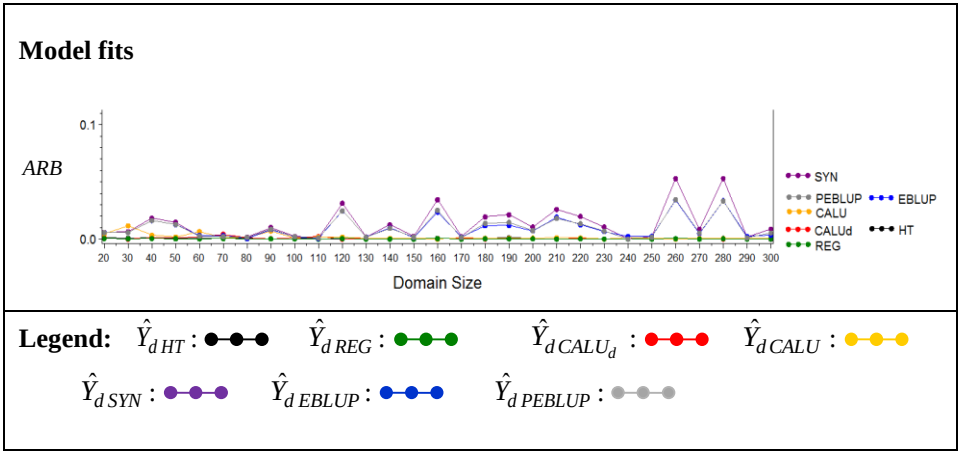


Figure 2b. Plots of the absolute relative bias of the estimators for sample size 464

Figures 3a and 3b display the coefficient of variation associated with the estimators. The coefficient of variation is reduced for all estimators except the HT estimator $\hat{Y}_{d\,HT}$ (which does not use any auxiliary information) and $\hat{Y}_{d\,CALU_d}$ (because the auxiliary variables for this estimator are equivalent in the two runs).

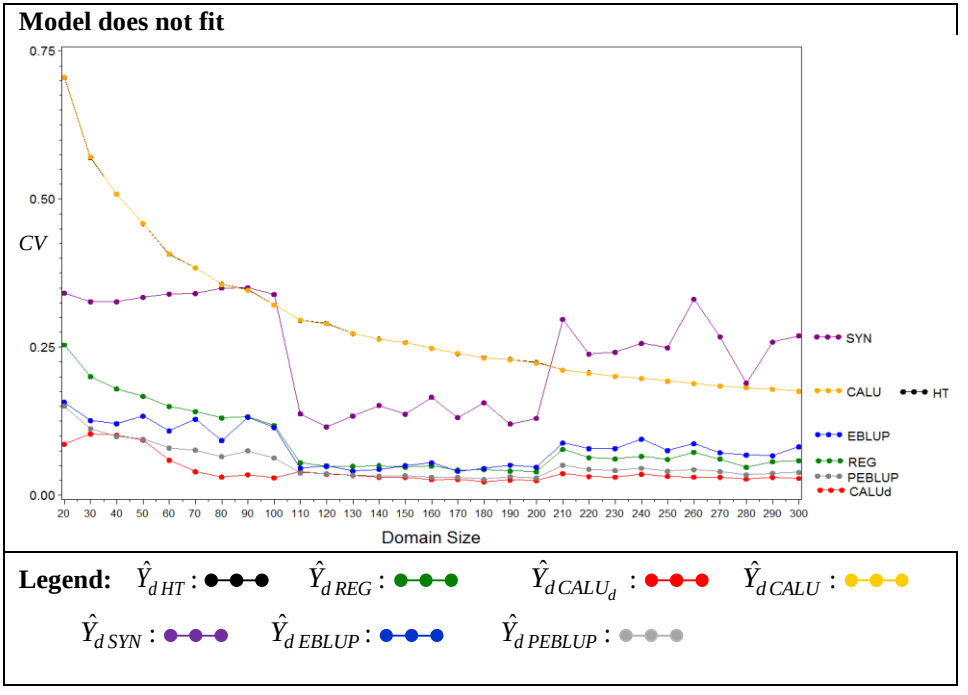


Figure 3a. Plots of the Coefficient of Variation of the Estimators for sample size 464

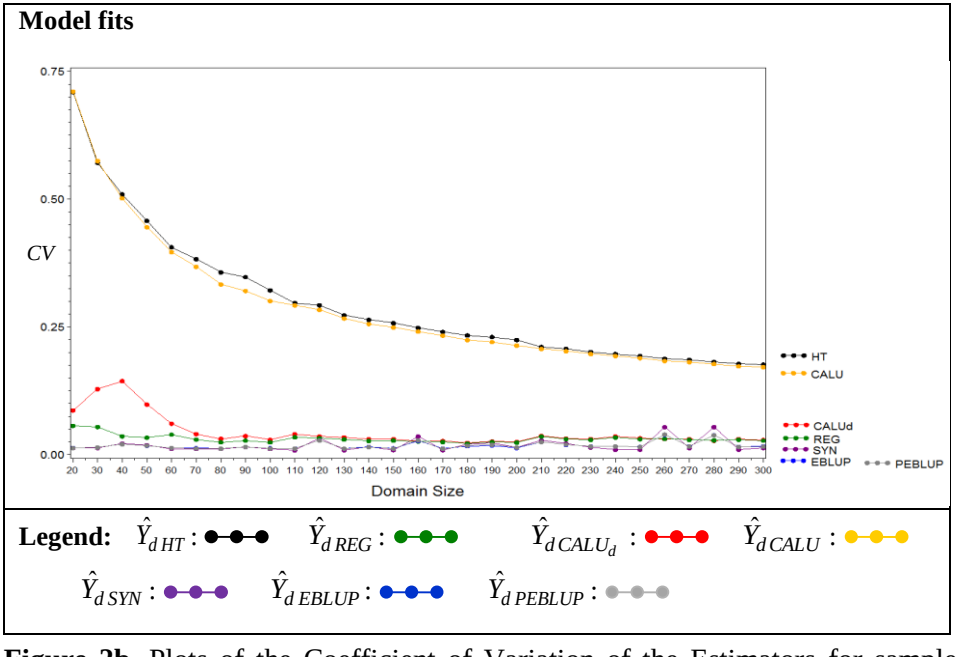


Figure 2b. Plots of the Coefficient of Variation of the Estimators for sample size 464

Figures 4a and 4b show a graphical display of the results for the average coefficient of variation $\overline{CV}(\hat{Y}_{EST})$ results given in Table 6. Under run 2, we see that $\hat{Y}_{d\,SYN}$, $\hat{Y}_{d\,EBLUP}$ and $\hat{Y}_{d\,PEBLUP}$ have the smallest $\overline{CV}(\hat{Y}_{EST})$. All three lines are indistinguishable as they are very close together. Under run 2, we see that $\hat{Y}_{d\,SYN}$, $\hat{Y}_{d\,EBLUP}$ and $\hat{Y}_{d\,PEBLUP}$ have the smallest $\overline{CV}(\hat{Y}_{EST})$. All three lines are indistinguishable as they are very close together.

Model does not fit

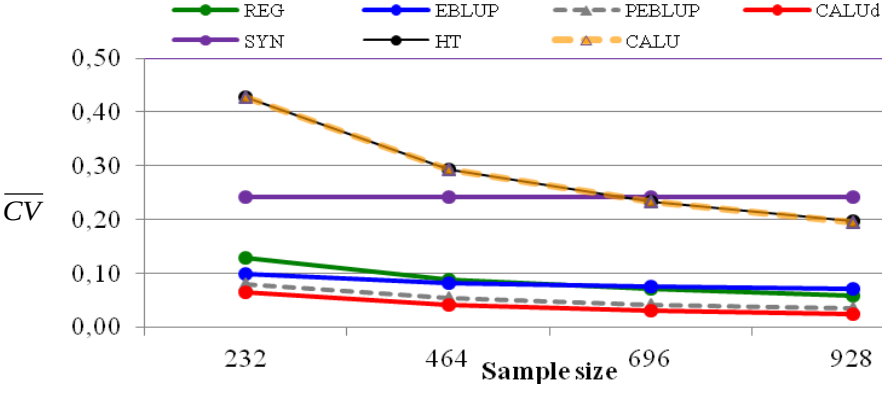


Figure 4a. Plots of the average coefficient of variation of the estimators by sample size

Model fits

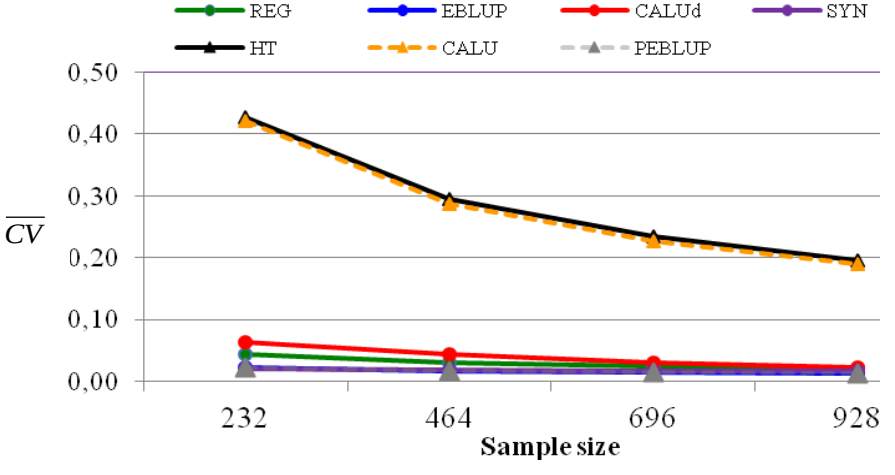


Figure 4b. Plots of the average coefficient of variation of the estimators by sample size

Figures 5a and 5b show a graphical display of the average relative efficiency of the estimators given in table 7. Under run 2, we note that $\hat{Y}_{d\,EBLUP}$ and $\hat{Y}_{d\,PEBLUP}$ have the highest $\overline{RE}(\hat{Y}_{EST})$ over the various sample sizes.

Model does not fit

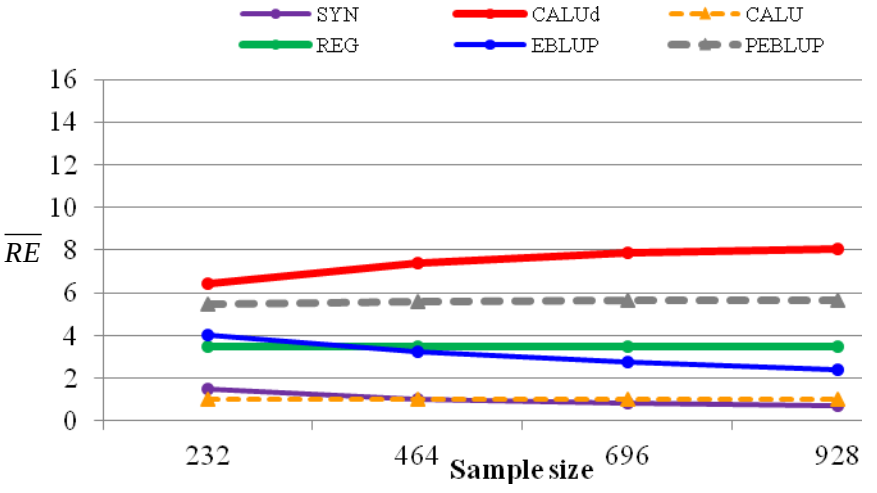


Figure 5a. Plots of the average relative efficiency of the estimators by sample size

Model fits

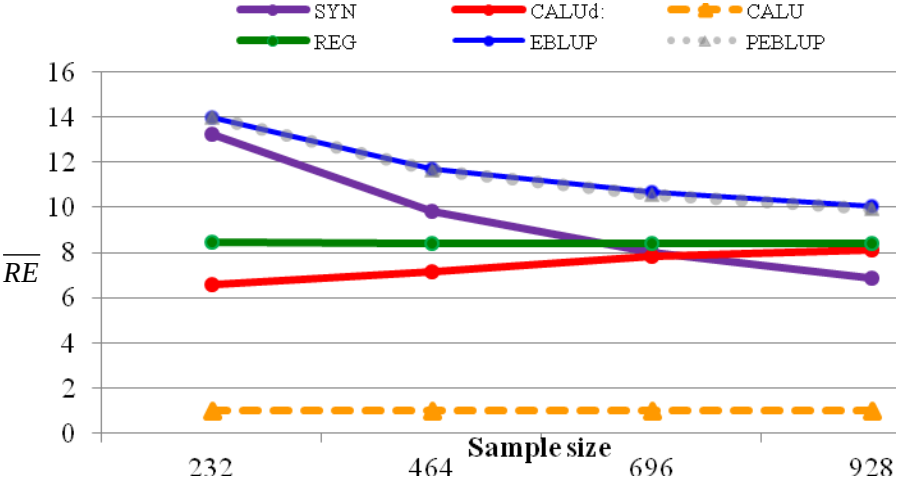


Figure 5b. Plots of the average relative efficiency of the estimators by sample size

4. Conclusions

We compared via simulation the behavior of a number of traditional domain and small area estimators. The sampling design used in the simulation, simple random sampling without replacement, is a simplification of the sampling design commonly used for business surveys (stratified simple random sampling without replacement). The estimators using the auxiliary data either reflected the model used to generate the population (model fits) or did not (model does not fit). The simulation design did not use unequal probability sampling. The additional complexity of using unequal probability sampling is that we would have had to modify our model-based small area estimators to account for possible informative sampling. However, since we used simple random sampling without replacement, we did not have to account for this problem.

The conclusions of our simulation are as follows. Comparing the efficiency between the traditional and small area estimators, the results very much depend on whether the model holds or not. The calibration estimator $\hat{Y}_{d\,CALU}$ which only uses auxiliary data at the population level is not efficient at the domain level whether the model holds or not. This is in contrast to $\hat{Y}_{d\,CALU_d}$ that uses auxiliary data at the domain level. The estimator $\hat{Y}_{d\,CALU_d}$ is the best traditional estimator to use when the model holds. Its average relative efficiency increases as the overall sample size increases. Its weakness is in the smaller domains, where the expected sample size is smaller than three units, as it cannot be defined when the auxiliary data consists of two auxiliary variables; in general, when there are p auxiliary variables, we are not be able to define $\hat{Y}_{d\,CALU_d}$ when the sample size is smaller than p+1 auxiliary variables. When the model does not hold, $\hat{Y}_{d\,REG}$ is the best traditional estimator to use. However, it is outperformed by the small area estimators $\hat{Y}_{d\,SYN}$, $\hat{Y}_{d\,EBLUP}$ and $\hat{Y}_{d\,PEBLUP}$. The small area estimator $\hat{Y}_{d\,EBLUP}$ is the most efficient one when the model holds, although it is closely followed by $\hat{Y}_{d\,SYN}$ and $\hat{Y}_{d\,PEBLUP}$. When the model does not hold, the $\hat{Y}_{d\,PEBLUP}$ estimator is the most efficient small area estimator; an explanation for this is that it is design-consistent.

Acknowledgement

We would like to thank the referee for his constructive comments that substantially improved this paper.

REFERENCES

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

DATTA, G. S., (2009). Model-based approach to small area estimation. *Handbook of Statistics*, 29, 251–288.

DEVILLE, J. C., SÄRNDAL, C. E., (1992). Calibration estimation in survey sampling. *Journal* of *the American Statistical Association*, 87(418), 376–382.

ESTEVAO, V., HIDIROGLOU, M. A., YOU, Y., (2014). Methodology Software Library - Small area Estimation Methodology Specifications for Area and Unit Level based Models. Technical Report, Statistics Canada.

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

LEHTONEN, R., VEIJANEN, A., (2009). Design-based methods of estimation for domains and small areas. *Handbook of statistics*, 29, 219–249.

PFEFFERMANN, D., (2013). New important developments in small area estimation. *Statistical Science*, 28(1), 40–68.

PFEFFERMANN, D., SVERCHKOV, M., (2007). Small Area Estimation Under Informative Probability Sampling of Areas and Within the Selected Areas. *Journal of the American Statistical Association* 102 (480), 1427–1439.

RAO, J. N. K., (2003). *Small Area Estimation:* John Wiley & Sons.

SINGH, A. C., MIAN, I. U. H., (1995). Generalized Sample Size Dependent Estimators for Small Areas, *Proceedings of the 1995 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 687–701.

SINGH, M. P., GAMBINO, J., MANTEL, H., (1994). Issues and Strategies for Small Area Data. *Survey Methodology*, 20 (1), 3–22.

WOODRUFF, R. S., (1966). Use of a Regression Technique to Produce Area Breakdowns of the Monthly National Estimates of Retail Trade. *Journal* of *the American Statistical Association*, 61 (314), 496–504.

YOU, Y., RAO, J. N. K., (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights, *Canadian Journal* of *Statistics*, 30, 431–439**.**

VERRET, F., RAO, J. N. K., VERRET, F., RAO, J. N. K., HIDIROGLOU, M. A., (2015). Model-based small area estimation under informative sampling. To appear in the December 2015 issue of *Survey Methodology*.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 155–158

ABOUT THE AUTHORS

Breidt, Jay is a Professor in the Department of Statistics, Colorado State University. He is a Fellow of the American Statistical Association, a Fellow of the Institute of Mathematical Statistics, and an elected member of the International Statistical Institute. His research interests include time series, nonparametric regression, environmental statistics, and design and estimation for complex surveys.

Burgard, Jan Pablo is an Assistant Professor of Statistics at the Economics Department of Trier University and a Principal Investigator at the ALOP research training group focusing on numerical optimization. His main research interests lie in small area estimation, survey design, and computational statistics, with focus on statistical modelling, multivariate statistics, and Monte Carlo methods. The main fields of application are register-based censuses, poverty indicators, the estimation of regional morbidity figures, and the measurement of biodiversity.

Chakraborty, Adrijo is a Statistician at NORC at the University of Chicago. He received his PhD in statistics from the University of Georgia in 2014. His research interests include small area estimation, Bayesian statistics, survey sampling, and public health.

Datta, Gauri Sankar is a Professor of Statistics at the University of Georgia and a Mathematical Statistician at the U.S. Census Bureau. Dr. Datta's research interests include small area estimation, Bayesian methods with applications to syndromic surveillance and the neuro-image analysis of fMRI data. He has published his methodological and applied research extensively in the leading journals of statistics. He is an elected fellow of the American Statistical Association and the Institute of Mathematical Statistics.

Erciulescu, Andreea is a Research Associate at the National Institute of Statistical Sciences (NISS), working on projects conducted by USDA National Agricultural Statistics Services (NASS). Erciulescu earned her Ph.D. in Statistics from the Department of Statistics, at Iowa State University. Her research interests include small area estimation, mixed models, resampling techniques, computational statistics, and survey statistics.

Fuller, Wayne A. is Emeritus Distinguished Professor in Statistics and Economics at Iowa State University. He has published in more than twenty journals and is the author of the texts *Introduction to Statistical Time Series*, *Measurement Error Models*, and *Sampling Statistics.* As a member of the Survey Group at Iowa State University, he had primary responsibility for developing estimation procedures for a large longitudinal national survey called the U.S. National Resources Inventory. His research interests in survey sampling include regression estimation, small area estimation, imputation, and multiple phase sampling. His research in time series concentrated on autoregressive processes, particularly those with a unit root.

Gabler, Siegfried is Team Leader of Statistics at GESIS - Leibniz-Institute for the Social Sciences. He is an elected member of the International Statistical Institute and member of the Sampling Expert Panel of the European Social Survey. He teaches as Privatdozent at the University of Mannheim. His research area covers sampling designs, especially for telephone surveys and cross-cultural surveys, weighting, design effects, and decision theoretic justification of sampling strategies.

Ganninger, Matthias is Senior Data Scientist at Roche Diagnostics. He received his PhD in statistics from the University of Trier in 2009. While working for GESIS - Leibniz-Institute for the Social Sciences he specialized in design effects and variance estimation. During his postdoctoral years, he focused on small area estimation, computational statistics, and Monte-Carlo simulation techniques. In 2013, he joined Elsevier Health Analytics where he specialized on modelling health insurance claims data. Since early 2015, Matthias Ganninger has been building data science capabilities at Roche Diagnostics on a global level.

Guadarrama, Maria is a PhD student at the Department of Statistics, Carlos III University of Madrid. Her main research domain is small area estimation under complex sampling designs. She is also interested in the estimation of general parameters in small areas, panel data analysis, and Bayesian statistics.

Hernandez-Stumpfhauser, Daniel is a Postdoctoral Research Associate at the Department of Biostatistics at the University of North Carolina at Chapel Hill, NC, USA. He has a PhD in Statistics from Colorado State University. His research interests include scalable Bayesian inference, variational Bayes, directional statistics, and Bayesian analysis of survey data.

Hidiroglou, Michael A. was Director of the Statistical Research and Methodology Division at Statistics Canada at the time that this article was written. He is currently Senior Research Advisor at the Business Survey Methods Division at Statistics Canada. Dr. Hidiroglou is a Fellow of the American Statistical Association and an elected member of the International Statistical Institute. His research interests are in the areas of sampling, data collection, and small area estimation.

Kolb, Jan-Philipp is Senior Statistician at GESIS - Leibniz-Institute for the Social Sciences. He received his PhD in statistics from the University of Trier in 2012. While working at the chair for economic and social statistics at the University of Trier he specialized in data analysis and the generation of synthetic universes as a basis for simulation. During his postdoctoral years, he focused on computational statistics, integration of geodata, and Monte-Carlo simulation techniques.

Kordos, Jan graduated from the Jagiellonian University and the University of Wroclaw (in mathematical statistics, 1955); PhD in Econometrics from the Academy of Economics, Katowice, Poland (1965), and Professorship (1990). He worked as the Chief of the Methodology Section at the Division of Living Conditions, Central Statistical Office/CSO (1955-1966) and of the Laboratory of Mathematical Methods at the Research Center of Statistics and Economics (CSO, 1966-74). He served as the FAO Adviser in Agricultural Statistics in Ethiopia (1974-80). He acted as Director of the Division Demographic and Social Surveys (1981-92) and as Vice President of the CSO Poland (1992-96). He was lecturing and training on agricultural Statistics in China in the late 1980s, and also in Kathmandu, Nepal (1991). During 1994-96 he served as the World Bank Consultant in Household Budget Surveys in Latvia and Lithuania. He was President of the Polish Statistical Association (1985-94). He was founder and editor-in-chief of *Statistics in Transition* (1993-2007). Now, he is Professor of Statistics at the Warsaw Management University and Adviser to the President of CSO. His publications include four books and over three hundreds articles and other papers. He is an elected member of the International Statistical Institute since 1974.

Mandal, Abhyuday is an Associate Professor in the Department of Statistics of the University of Georgia, USA. He received his Ph.D. from the Georgia Institute of Technology in 2005. His research interests include small area estimation and design of experiments.

Molina, Isabel is an Associate Professor at the Department of Statistics, Carlos III University of Madrid. She is an elected member of the International Statistical Institute. She has published over 30 research papers including several book chapters and coauthored the Wiley book "Small area estimation, second edition". Her research interests include small area estimation, mixed models, robust methods, and resampling techniques (bootstrap).

Münnich, Ralf is a Full Professor and the Head of the Economic and Social Statistics research group at Trier University. He has led several large-scale European research projects and the German Census 2011 sampling and estimation project. Currently, he is heading the RIFOSS research initiative and is a principal investigator in the ALOP research training group focusing on discrete optimization in survey statistics. Ralf Münnich is an elected member of the International Statistical Institute, Member of the Board of the German Statistical Society, and editor-in-chief of AStA Wirtschafts- und Sozialstatistisches Archiv. His main research interests focus on survey sampling, variance estimation, small area estimation, and Monte-Carlo and microsimulation methods.

Opsomer, Jean D. is Professor and Chair in the Department of Statistics, Colorado State University. He is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics, and is an elected member of the International Statistical Institute. His main research interests are survey statistics, nonparametric methods, and environmental statistics.

Rao, J. N. K. is a Distinguished Research Professor in the School of Mathematics and Statistics, Carleton University, Ottawa, Canada. He is also a consultant to Statistics Canada on sample survey methodology. His main research interests are in survey sampling theory and methods. He has published numerous research papers as well as the widely cited book "Small Area Estimation" (Wiley 2003) and a second edition of this book jointly with Isabel Molina (Wiley 2015). He is an editorial advisor for the Wiley series in Survey Methodology, and currently sits on the editorial board of the Survey Methodology journal. He has served on the Advisory Committee for Statistical Methodology of Statistics Canada since 1985. The professional honours he has received include Honorary Doctorates from the University of Waterloo, Canada (2008) and Catholic University of the Sacred Heart, Italy (2013), Waksberg Award for Survey Methodology (2005) and 1993 Gold Medal of the Statistical Society of Canada in recognition of "fundamental research achievements in the theory and practice of surveys". He is a Fellow of the Royal Society of Canada, American Statistical Association and Institute of Mathematical Statistics. He delivered the prestigious Annual Morris Hansen Lecture in 1998.

E R R A T U M

In the previous joint issue of the *Statistics in Transition new series* and *Survey Methodology* after the name of a Guest Editor, Professor Risto Lehtonen, mistakenly was given his affiliation ("University of Jyväskylä"). The proper information should read as follows: Risto Lehtonen, University of Helsinki. We apologize to Professor Risto Lehtonen and to the readers for this mistake.