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 1

CONTENTS

From the Editors 483

From the Guest Editors (Part 1) 485

Submission information for authors 489

J. N. K. Rao, Inferential issues in model-based small area estimation: some new developments 491

Daniel Bonnéry, Yang Cheng, Neung Soo Ha, Partha Lahiri, Triple-goal estimation of unemployment rates for U.S. states using the U.S. Current Population Survey data 511

Jan A. van den Brakel, Bart Buelens, Covariate selection for small area estimation in repeated sample surveys 523

Forough Karlberg, Small area estimation for skewed data in the presence of zeroes 541

Carolina Franco,William R. Bell, Borrowing information over time in binomial/logit normal models for small area estimation 563

Angela Luna, Li-Chun Zhang, Alison Whitworth, Kirsten Piller, Small area estimates of the population distribution by ethnic group in England: a proposal using structure preserving estimators 585

Jan Pablo Burgard, Ralf Münnich - SAE teaching using simulations 603

Elżbieta Gołata, SAE education challenges to academics and NSI 611

About the Authors 631

Acknowledgements to reviewers 635

Index of Authors 639

Volume 16, Number 4, December 2015

Editorial Office **ISSN 1234-7655**

Marek Cierpiał-Wolan, Ph.D.: Scientific Secretary
m.wolan@stat.gov.pl
Beata Witek: Secretary
b.witek@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. 16, No. 4, pp. 483–484

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 SEA 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 are split into two parts: the first one appears in this thematic issue, while the second one will be published in the March 2016 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.SA. 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. 16, No. 4, pp. 485–488

FROM THE GUEST EDITORS (PART 1)

The first part of this Joint Issue of Statistics in Transition and Survey Methodology includes eight 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 six contributions in this thematic issue of SIT and SMJ represent articles that are firmly methodological in their perspective. The first paper, by J.N.K. Rao provides a unifying perspective for the remaining five contributions. In this review paper, Rao highlights important new developments in SAE since the publication of his encyclopedic 2003 book. As he notes in his abstract, much of this new methodological development has focused on addressing the practical issues that arise when model-based SAE methods are applied in practice. An important dichotomy in this regard follows from the nature of the available data for SAE. Historically, such data have been area level aggregates of one form or another, typically direct sample-based estimates. Issues addressed in Rao's paper then include the choice of appropriate weights for these aggregates as well as methods for dealing with the not uncommon situation where there is a negligible area level variance component in the basic area-level model (the so-called Fay-Herriot model) used to smooth these aggregates across the areas, or where this smoothing model is necessarily non-linear, reflecting a GLM for the underlying survey variable. Issues associated with estimation of both unconditional as well as conditional MSEs of these model-based estimators are also discussed. In the second half of his paper, Rao switches his attention to SAE where unit level data from the small areas of interest are available. This is a fast-growing set of applications, reflecting new capabilities in data collection. Here, the focus is on sample weighting and benchmarking as important requirements for users interested in design consistency of SAE outputs, together with important new developments in dealing with outliers in the survey data, applications to poverty mapping and dealing with informative sampling methods. Model selection and checking is extremely important in the unit level case, and the paper briefly describes some new developments in this regard.

The next three papers in this issue focus on a new methodology for area level SAE. The first, by Bonnery, Cheng, Ha and Lahiri, notes that users of SAE outputs typically require more than just estimates of area averages, and are often interested in small area distributions as well as rankings across small areas. In this context, these authors develop a triple goal SAE methodology for US state level unemployment, with estimates structured so that they are simultaneously efficient for estimation of area level average unemployment as well as the empirical distribution of area level unemployment, while also staying as close as possible to the actual ranking of the real small area means. An interesting idea that is discussed in this paper is the fact that in practice it is not just one area average that is of interest, but an "ensemble" of such averages corresponding to the area-level distribution of a characteristic of interest. This immediately leads to a corresponding ensemble of models, which these authors fit using a Bayesian MCMC approach.

The general theme of the usefulness of incorporating time series information in SAE solution is repeated in the paper by van den Brakel and Buelens. Here, though the attention is directed towards appropriate model specification when the estimation must be carried out at regular intervals, using data from repeated surveys and practical considerations rule out survey-specific model optimisation. An approach to covariate selection for small area survey estimates obtained from a repeated survey under a Fay-Herriot specification is defined, with the model specification carried out simultaneously over a number of "editions" of the survey while being constrained to be the same for each edition. The final model is chosen by minimising the average conditional AIC over all the editions, with the small area estimates at each time period computed using a Hierarchical Bayes approach.

The next paper, by Karlberg, switches gears and considers SAE under a unit level model. In particular, in this paper Karlberg addresses two of the difficult issues that arise when the available unit level data are non-negative values drawn from an economic population, as would be the case for a business survey. These conditions often lead to a highly right-skewed distribution of the sample data values, with outliers a not uncommon feature, together with the presence of excess zeros. Both of these data characteristics are not conducive to SAE based on the industry standard linear mixed model for unit level data. Instead, Karlberg combines a log scale linear mixed model for the strictly positive data (to deal with their high skewness) and a logistic model for the presence of zero values (a hurdle model) in order to define a specification for the zero-inflated observed data. Simulation results for SAE based on this approach are promising, but application to a real business survey data set turns out to be disappointing, reflecting the very complex nature of such data. Clearly further research is needed for SAE in business surveys.

The fifth paper, by Franco and Bell, shows how the Fay-Herriot approach can be extended to where the underlying averages are derived from binary survey variables, so that the basic area-level model can be specified as linear on a logit scale. This model is then combined with time series of aggregates from the small areas, allowing for information to be "borrowed" across both time and space. An application to improving county-level poverty estimates in the SAIPE programme of the US Bureau of the Census is used to demonstrate the efficiency gains of the approach.

The sixth paper, by Luna, Zhang, Whitworth and Piller, represents a fundamental departure from the random area effect-based SAE models that underpin the previous papers. Here, the underlying data consist of historical counts, represented by an out-of-date census (or register)-based cross-tabulation of interest, where one of the dimensions of the tabulation is the area identifier, as well as up-to-date information on margins of the cross-tabulation derived from a current survey. Such data are naturally modelled using a log-linear specification, and the authors consider the use of a generalized SPREE approach to recover the current cross-tabulation. Alternative GSPREE models with increasingly complex interaction structure are investigated and applied to estimation of population counts within ethnic group in small areas in the United Kingdom. Interestingly, these authors report that for these data more complex model specifications do not necessarily lead to improvement in the resulting survey estimates, essentially because the sparse nature of the available data does not allow these more complex models to be adequately fitted.

The last two contributions focus on small area education. Small area estimation is gaining increasing popularity among survey statisticians, economists, sociologists and many others. Unfortunately, small area courses are offered only in a handful of universities and that too just as an elective. However, there is a definite need for small area teaching, and the papers by Burgard and Münnich as well as Golata have addressed this very important issue. The paper by Burgard and Münnich has hit the mark very directly. What the paper emphasizes is that rather than giving a series of lectures on the different small area techniques and the associated theory behind them, it is more important to combine the theory with actual simulations. In this way, students can have hands on experience of the subject as well as are able to make a comparison of the different small area methods which they have learnt. Like Burgard and Münnich, Golata also appreciates very well the need for small area education. To this end, she conducted a survey with participants from both the academics and National Statistical Institutes. Her objective went beyond questions on small area teaching, and enquired several related pertinent questions such as risks encountered in applying SAE as well as important sources on SAE developments. The results of her survey are listed in a series of tables and graphs to provide the reader with a better understanding of the state of the art.

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 would like to thank all the authors for taking the time to turn their SAE 2014 presentations into the interesting and thought provoking papers published here. We acknowledge the efforts of Giovanna Ranalli, Nicola Salvati, Hukum Chandra and Timo Schmid, who helped review the first six papers: their encouraging and productive comments directly contributed to their obvious quality.

Raymond Chambers and **Malay Ghosh**
Guest Editors

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 489−490

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. 16, No. 4, pp. 491–510

INFERENTIAL ISSUES IN MODEL-BASED SMALL AREA ESTIMATION: SOME NEW DEVELOPMENTS

J. N. K. Rao[1]

ABSTRACT

Small area estimation (SAE) has seen a rapid growth over the past 10 years or so. Earlier work is covered in the author's book (Rao 2003). The main purpose of this paper is to highlight some new developments in model-based SAE since the publication of the author's book. A large part of the new theory addressed practical issues associated with the model-based approach, and we present some of those methods for area level and unit level models. We also briefly mention some new work on synthetic estimation of area means or totals based on implicit models.

Key words: area level models, complex parameters, informative sampling, model misspecification, robust estimation, unit level models.

1. Introduction

The author's 2003 Wiley book (Rao 2003) provided a comprehensive account of the theory and methods of model-based small area estimation (SAE), which borrows strength through explicit models linking related small areas. Model-based SAE, both in theory and applications, has seen rapid growth over the past 10 years due to growing demand for reliable small area statistics. In a review paper, Pfeffermann (2013) says “The diversity of new problems investigated is overwhelming, and the solutions proposed are not only elegant and innovative, but also very practical”.

The main purpose of this paper is to highlight some new developments in model-based SAE since the publication of the author's 2003 book. A large part of the new theory addressed practical issues associated with the model-based approach, and we present some of those methods for area level and unit level models. We also briefly mention some new work on synthetic estimation of area means or totals based on implicit models.

[1] Carleton University, Ottawa, Canada. E-mail:jrao34@rogers.com.

2. Synthetic estimation based on weight sharing

Let Y_i be the total of a variable of interest y for domain (or area) i. Let s be a probability sample from a finite population with associated inclusion probabilities π_k and values $y_k, k \in s$. Then, a basic area-specific direct estimator of Y_i is given by the expansion estimator

$$\hat{Y}_i = \sum\nolimits_{k \in s(i)} w_k y_k , \tag{2.1}$$

where $s(i)$ is the subsample of units belonging to area i and $w_k = 1/\pi_k$.

Improved direct estimators (such as generalized regression estimators) can also be obtained using supplementary population information. Such direct area estimators are not useful or feasible for SAE if area-specific samples of inadequate sizes or no samples are available.

We first present synthetic estimation of small area totals based on weight sharing. The basic idea behind weight sharing is to produce weights w_{ij} for each area i and each unit $j \in s$ that satisfy the calibration property

$$\sum_{j \in s} w_{ij} x_j = X_i, \quad i = 1,...,m \tag{2.2}$$

and the weight-sharing property

$$\sum_{i=1}^{m} w_{ij} = w_j, \qquad j \in s \tag{2.3}$$

where X_i is the known area total of an auxiliary vector variable x. The weight-sharing (WS) synthetic estimator of the area total Y_i is given by

$$\hat{Y}_{iWS} = \sum_{j \in s} w_{ij} y_j . \tag{2.4}$$

The weight-sharing property ensures that the associated estimators $\hat{Y}_{iWS}$ add up to the direct estimator $\hat{Y} = \sum_{j \in s} w_j y_j$ of the population total $Y = \sum_{i=1}^{m} Y_i$, and the calibration property improves the efficiency of the estimator. The use of the same weight, w_{ij}, for all variables of interest used as y to produce small area estimates is of practical interest, particularly in micro-simulation modelling that can involve a large number of variables of interest. The estimator $\hat{Y}_{iWS}$ borrows strength from other areas because it makes use of all the sample values $y_j, j \in s$.

Schirm and Zaslavsky (1997) proposed an iterative method of finding the weights w_{ij} that satisfy (2.2) and (2.3), but it uses a model on the weights w_{ij} of the form $w_{ij} = \exp(x_j^T \beta_i + \delta_j)$, where β_i and δ_j are unknown coefficients.

Randrianasolo and Tille (2013) avoid modelling the weights w_{ij} by minimizing an information distance measure between the weights w_{ij} and w_j subject to the constraints (2.2) and (2.3), separately for each i. They used a two-step iteration by letting $w_{ij} = w_j q_{ij}$ such that the fractions q_{ij} satisfy $\sum_{i=1}^{m} q_{ij} = 1$ for each $j \in s$.

3. Basic area-level model

3.1. The model

Let $\bar{Y}_i$ be the mean of area i and $\hat{\bar{Y}}_i$ be a direct estimator of $\bar{Y}_i$. Poverty rate P_i is a special case of $\bar{Y}_i$ by letting $y = 1$ if the welfare variable for a household is below a specified poverty line and $y = 0$ otherwise. Estimation of poverty rates for small areas, such as municipalities, has received considerable attention worldwide in recent years. Data consists of direct estimators $\hat{\bar{Y}}_i$ and associated vectors of area-level covariates z_i for the m areas. Basic area-level model (also called Fay-Herriot (FH) model) consists of a linking model

$$\theta_i = g(\bar{Y}_i) = z_i^T \beta + v_i \ , v_i \sim_{\text{iid}} \text{N}(0, \sigma_v^2) \,, \tag{3.1.}$$

and a “matching” sampling model

$$\hat{\theta}_i = \theta_i + e_i, \ e_i \sim_{\text{ind}} N(0, \psi_i) \,, \tag{3.2}$$

where e_i is the sampling error with known variance ψ_i and independent of v_i (Fay and Herriot 1979). If all the areas in the population are not sampled, we assume that the model holds for the sampled areas $i = 1, ..., m$. We do not consider informative sampling of areas which causes sample selection bias and the model, assumed for all the population areas, may not hold for the sample.

Limitations of the FH model include the assumptions of known sampling variances ψ_i and zero mean sampling errors e_i.The latter assumption may not hold for non-linear functions $g(.)$ even approximately if the area sample size is small. An unmatched sampling model of the form $\hat{\bar{Y}}_i = \bar{Y}_i + h_i$ with zero mean sampling errors h_i avoids the latter difficulty with the sampling model (3.2).

Main advantages of the FH model are that it takes account of the sampling design through the model (3.2) on direct estimators and that it requires only area level covariates which are more easily available than unit level covariates. Current applications of the FH model include the estimation of the number of school age children in poverty in the US counties and school districts (Luery 2011) and the estimation of household poverty rates for the Chilean Communas (Casas-Cordero, Encina and Lahiri 2014). In the first application, $\theta_i = \log(Y_i)$ and the direct county estimates $\hat{Y}_i$ of area totals Y_i are obtained from the American Community Survey. In the second application, $\theta_i = \sin^{-1}\sqrt{P_i}$ and the direct estimates $\widehat{P}_i$ are obtained from a cross-sectional multi-purpose household survey. Excellent area-level covariates, based on administrative sources, are available in both applications.

3.2. "Optimal" estimation

For known parameters β and σ_v^2, the "best" predictor (BP) of θ_i under normality of the model errors v_i and the sampling errors e_i is given by

$$\tilde{\theta}_i^B = E(\theta_i \mid \hat{\theta}_i) = \gamma_i \hat{\theta}_i + (1-\gamma_i) z_i^T \beta \,, \tag{3.3}$$

where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \psi_i)$. The estimator $\tilde{\theta}_i^B$ is model unbiased for θ_i in the sense that $E(\tilde{\theta}_i^B - \theta_i) = 0$. It follows from (3.3.) that more weight is given to the direct estimator $\hat{\theta}_i$ if the model variance σ_v^2 is large relative to the sampling variance ψ_i, and more weight given to the synthetic estimator $z_i^T \beta$ if the sampling variance ψ_i is large. The mean squared error (MSE) of $\tilde{\theta}_i^B$ under the FH model is given by

$$MSE(\tilde{\theta}_i^B) = E(\tilde{\theta}_i^B - \theta_i)^2 = g_{1i}(\sigma_v^2) = \gamma_i \psi_i \,, \tag{3.4}$$

which shows that $\tilde{\theta}_i^B$ is significantly more efficient than the direct estimator $\hat{\theta}_i$ if γ_i is small. The estimator $g^{-1}(\tilde{\theta}_i^B)$, obtained by back transformation, is commonly used to estimate the area mean $\bar{Y}_i$. It is not optimal and also leads to model bias. In the Chilean application (Casas Cordero et al. 2014), the estimator of poverty rate P_i is given by $\sin^2 \tilde{\theta}_i^B$.

In practice, we replace (β, σ_v^2) in (3.3) by maximum likelihood (ML) or restricted ML (REML) estimators to get the empirical best (EB) predictor $\hat{\theta}_i^{EB}$ of θ_i. An empirical best linear unbiased predictor (EBLUP) without normality assumption, denoted by $\hat{\theta}_i^H$, has the same form as $\hat{\theta}_i^{EB}$, where the estimators of model parameters are obtained by a method of moments, see Rao (2003, Chapter 7) for details. We denote the estimators of model parameters by $(\hat{\beta}, \hat{\sigma}_v^2)$. The above methods of estimating σ_v^2 can lead to $\hat{\sigma}_v^2 = 0$. A drawback of using zero estimate of σ_v^2 is that the resulting EB estimate $\hat{\theta}_i^{EB}$ will attach zero weight to all the direct estimates $\hat{\theta}_i$ regardless of the area sample sizes. Giving a zero weight to the direct estimates for areas with large enough sample sizes is not appealing to the user, and substantial disagreement between EB and direct estimates can occur due to over shrinkage induced by the zero estimate of σ_v^2. This problem attracted considerable attention in the recent literature, leading to alternative methods of estimating model parameters that avoid a zero value for $\hat{\sigma}_v^2$. Methods studied include data-based truncation (Wang and Fuller 2003) and maximizing an adjusted likelihood function (Li and Lahiri 2010 and Yoshimori and Lahiri 2014).

Simulation results suggest that the EB estimator $\hat{\theta}_i^{YL}$, based on the Yoshimori and Lahiri (YL) estimator of σ_v^2, performs better in terms of MSE than the EB estimator $\hat{\theta}_i^{LL}$ based on the Li and Lahiri (LL) estimator of σ_v^2.

3.3. MSE estimation

3.3.1. Unconditional MSE

A difficulty with the EB estimator $\hat{\theta}_i^{EB}$ is that no closed-form expression for its MSE is available except for a few special cases. This difficulty has attracted a lot of attention in the SAE literature, leading to second-order approximations to MSE($\hat{\theta}_i^{EB}$) which in turn are used to derive second-order unbiased estimators of MSE. In particular, in the case of REML estimators of model parameters, a second order unbiased MSE estimator is given by

$$\text{mse}(\hat{\theta}_i^{EB}) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2), \qquad (3.5)$$

where the leading term $g_{1i}(\hat{\sigma}_v^2)$ is given by (3.3) with σ_v^2 replaced by $\hat{\sigma}_v^2$ and the remaining two terms in (3.5) are of lower order and account for the estimation of β and σ_v^2 respectively (see Rao 2003, section 7.1.5 for details). The MSE estimator of $\hat{\theta}_i^{YL}$ is obtained from (3.5) by substituting the YL estimator of σ_v^2 for $\hat{\sigma}_v^2$.The two MSE estimators are second –order unbiased in the sense that the bias is of lower order than $1/m$ for m large.

If σ_v^2 is suspected to be small relative to sampling variances ψ_i, then it could result in either a zero value or a very small value of $\hat{\sigma}_v^2$. In such cases, the second order unbiased MSE estimator (3.5) may lead to severe overestimation. An alternative is to conduct a preliminary test of the null hypothesis $\sigma_v^2 = 0$ at a reasonable test level, say 0.2, and then use the following MIX estimator of MSE($\hat{\theta}_i^{EB}$): $g_{2i}(0)$ if the null hypothesis is not rejected or $\hat{\sigma}_v^2 = 0$, otherwise use $\text{mse}(\hat{\theta}_i^{EB})$ given by (3.5). Similarly, a MIX estimator of MSE($\hat{\theta}_i^{YL}$) uses $g_{2i}(\hat{\sigma}_{v,YL}^2)$ if the null hypothesis is not rejected, otherwise $\text{mse}\left(\hat{\theta}_i^{YL}\right)$. Simulation studies suggest that the MIX estimators perform better than the second order unbiased estimators in terms of relative bias when σ_v^2 is small (Molina, Rao and Datta 2015).

The analytical approximation (3.5) based on linearization is valid for the EB estimator $\hat{\theta}_i^{EB}$, but not readily extendable to MSE estimation for the estimator of area mean given by $g^{-1}(\hat{\theta}_i^{EB})$. On the other hand, parametric bootstrap is readily applicable to general estimators. We describe the method for estimating MSE($\hat{\theta}_i^{EB}$), but the method follows along the same lines for estimating the MSE of general estimators. Assuming normality of v_i and e_i and $\hat{\sigma}_v^2 > 0$, we generate a bootstrap sample $\{((\hat{\theta}_{i*}, z_i), i = 1, ..., m\}$ in two steps: (1) Generate θ_{i*} from $N(z_i^T \hat{\beta}, \hat{\sigma}_v^2)$ independently for $i = 1, ..., m$. (2) Generate $\hat{\theta}_{i*}$ from $N(\theta_{i*}, \psi_i)$.

From the bootstrap data $\{(\hat{\theta}_{i*}, z_i), i - 1, ...m\}$ compute the estimate $\hat{\theta}_{i*}^{EB}$ in the same manner as $\hat{\theta}_i^{EB}$ computed from the sample data $\{(\hat{\theta}_i, z_i), i = 1, ..., m\}$. Repeat the above steps a large number, B, of times to get B bootstrap EB estimates $\hat{\theta}_{i*}^{EB}(1), ..., \hat{\theta}_{i*}^{EB}(B)$ and the bootstrap values of θ_i, denoted by $\theta_{i*}(1), ..., \theta_{i*}(B)$. A bootstrap MSE estimator is then given by

$$\text{mse}_B(\hat{\theta}_i^{EB}) = B^{-1} \sum_{b=1}^{B} [\hat{\theta}_{i*}^{EB}(b) - \theta_{i*}(b)]^2 . \qquad (3.6)$$

Noting that the bootstrap FH model is a replica of the FH model for the sample data, it follows that $\text{mse}_B(\hat{\theta}_i^{EB}) \approx g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + g_{3i}(\hat{\sigma}_v^2)$.

Comparing this approximation to (3.5) it follows that the bootstrap MSE estimator is not second order unbiased. It is possible to obtain second order unbiased bootstrap MSE estimators by generating second phase bootstrap samples from each first phase bootstrap sample (Hall and Maiti 2006).

3.3.2. Conditional MSE

In the previous subsection we presented some results on estimating the unconditional MSE of the EB estimator $\hat{\theta}_i^{EB}$. However, it is more appealing to consider the estimation of conditional MSE of $\hat{\theta}_i^{EB}$, treating the small area parameters θ_i as fixed unknown parameters. The conditional MSE is given by $\mathrm{MSE}_p(\hat{\theta}_i^{EB}) = E[(\hat{\theta}_i^{EB} - \theta_i)^2 \mid \theta]$, where $\theta = (\theta_1, ..., \theta_m)^T$.

Expressing $\hat{\theta}_i^{EB}$ as $\hat{\theta}_i^{EB} = \hat{\theta}_i + h_i(\hat{\theta})$, where $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)^T$ and $h_i(\hat{\theta}) = -(1-\hat{\gamma}_i)(\hat{\theta}_i - z_i^T\hat{\beta})$, an exactly unbiased estimator of conditional MSE is given by

$$\mathrm{mse}_p(\hat{\theta}_i^{EB}) = \psi_i + 2\psi_i[\partial h_i(\hat{\theta}) / \partial\hat{\theta}_i] + h_i^2(\widehat{\theta}). \qquad (3.7)$$

Datta, Kubokawa, Molina and Rao (2011a) gave an explicit expression for the derivative in the second term of (3.7) when REML estimators of model parameters are used.

The conditional MSE estimator (3.7) can take negative values and it can be highly unstable. Datta et al. (2011a) conducted a small simulation study under the conditional set-up for $m = 30$ and found that its coefficient of variation (CV) can be very high (ranged from 13% to 393%), especially for areas with large sampling variances ψ_i. Therefore, the conditional MSE estimator is not reliable as the estimator of the conditional MSE, although conditionally unbiased. It would be worthwhile to study if the bootstrap MSE estimator (3.6) can track the conditional MSE and still perform well in terms of CV.

3.4. Parametric bootstrap confidence intervals

Bootstrap data $\{(\hat{\theta}_{i*}, z_i), i = 1, ..., m\}$ can be used to construct confidence intervals on θ_i. Chatterjee, Lee and Lahiri (2008) proposed to use the bootstrap data to approximate the distribution of the pivotal $t_i = (\hat{\theta}_i^{EB} - \theta_i) / [g_{1i}(\hat{\sigma}_v^2)]^{1/2}$. The bootstrap value of t_i is given by $t_i^{\backslash *} = (\hat{\theta}_{i*}^{EB} - \hat{\theta}_i^{EB}) / [g_{1i}(\hat{\sigma}_{v*}^2)]^{1/2}$. In practice, we generate a large number, B, of bootstrap pivotals, denoted by $t_i^*(1), ..., t_i^*(B)$, and determine the lower and upper points, q_1 and q_2 such that the area between the lower and upper points of the empirical bootstrap distribution is equal to a specified nominal level $1-\alpha$. A bootstrap $(1-\alpha)-$ level interval on θ_i is then obtained from $q_1 \le t_i \le q_2$ as

$$I_i^{CLL}(\alpha) = [\hat{\theta}_i^{EB} - q_2\{g_{1i}(\hat{\sigma}_v^2)\}^{1/2}, \hat{\theta}_i^{EB} - q_1\{g_{1i}(\hat{\sigma}_v^2)\}^{1/2}] =: (c_{1i}, c_{2i}) \quad (3.8)$$

Chatterjee et al. (2008) showed that, under regularity conditions and normality of v_i and e_i , the interval (3.8) is second order correct in the sense that the error in its coverage is lower order than m^{-1}. The corresponding $(1-\alpha)-$ level second order correct bootstrap interval on the mean $\bar{Y}_i$ is obtained by back transformation as $[g^{-1}(c_{1i}), g^{-1}(c_{2i})]$, provided $\theta_i = g(\bar{Y}_i)$ is a one-to-one function.

Casas-Cordero et al. (2014) used bootstrap intervals for the poverty rates P_i in Chilean Communas. In their case, the bootstrap confidence interval on the poverty rate P_i is given by $[\sin^2(c_{1i}), \sin^2(c_{2i})]$.

3.5. Practical issues

We need to address several practical issues in implementing EB estimation under the FH model. Those issues include (i) covariates subject to sampling or measurement errors, (ii) unknown sampling variances ψ_i, (iii) linking model (3.2) incorrectly specified and (iv) benchmarking EB estimators to a reliable direct estimator at an aggregate level. We give a brief account of methods proposed to deal with the above practical issues.

Covariates subject to sampling errors. The FH model assumes that the covariates z_i are population values not subject to sampling or measurement errors. However, some of the covariates might be obtained from an independent survey with much larger area sample sizes than the survey of interest. For example, Ybarra and Lohr (2008) studied the estimation of mean body mass index θ_i for 50 small areas in the US using direct estimates $\hat{\theta}_i$ obtained from the 2003-2004 U. S. National Health and Nutrition Examination Survey (NHANES); NHANES values are obtained through medical examinations. They also used direct estimates $\hat{z}_i$ of the mean self-reported body mass index z_i, obtained from the 2003 U. S. National Health Interview Survey (NHS), as the covariate in the FH model. Area sample sizes for the NHANES are much smaller than those for the NHS and the direct estimates $\hat{z}_i$ are reliable and strongly correlated with the direct estimates $\hat{\theta}_i$. Ybarra and Lohr (2008) derived an optimal estimator of θ_i under the above set-up assuming that the variance of $\hat{z}_i$ is known. This estimator has the same form as the naïve estimator $\hat{\theta}_i^{EB}$ with z_i replaced by $\hat{z}_i$, but it attaches a larger weight to the direct estimator than the naïve estimator. The proposed estimator can lead to substantial gain in efficiency over the naïve estimator under the above set-up. Also, unlike the naïve estimator, it is never less efficient than the direct estimator. Marchetti et al. (2015) applied the Ybarra-Lohr estimator to estimate poverty rates in Tuscany region of Italy, using $\hat{z}_i$ derived from "big data" on mobility comprised of different car journeys automatically tracked with a GPS device. We predict that the use of big data will receive considerable attention in future SAE applications.

Unknown sampling variances. The FH model assumes known sampling variances ψ_i. Wang and Fuller (2003) and Rivest and Vandal (2003) relaxed this assumption by substituting a direct estimator $\hat{\psi}_i$ based on unit level data, for the case of $\theta_i = \overline{Y}_i$. The effect of estimating the sampling variances is to inflate the MSE of the EB estimator relative to the case of known sampling variances. As a result, the MSE estimator (3.5) with $\hat{\psi}_i$ substitute for ψ_i is no longer second order unbiased and it could lead to significant underestimation of the true MSE.

The above authors derived second order unbiased MSE estimators that contain an extra term arising from the estimation of ψ_i. On the other hand, if "smoothed" estimates $\hat{\psi}_{iS}$ of the sampling variances are used in the EB estimator, then no adjustment to the MSE estimator (3.5) is needed, provided the number of areas, m, is not small (Rivest and Vandal 2003).

Incorrectly specified linking model. The EB estimator uses the assumed linking model to estimate the model parameters β and σ_v^2. Jiang, Nguyen and J. S. Rao (2011) suggested an alternative approach that does not appeal to the linking model to estimate the model parameters and uses only the sampling model (3.1). They minimize the total sampling MSE of the best estimators $\tilde{\theta}^B = (\tilde{\theta}_1^B, ..., \tilde{\theta}_m^B)^T$ with respect to the model parameters. The total MSE is given by $E_p(|\tilde{\theta}^B - \theta|^2) = \sum_{i=1}^m E_p(\tilde{\theta}_i^B - \theta_i)^2$, where E_p denotes the expectation with respect to the sampling model conditional on $\theta = (\theta_1, ..., \theta_m)^T$. The resulting estimators of β and σ_v^2, called Best Predictive Estimators (BPEs), are then substituted into $\tilde{\theta}_i^B$ to get Observed Best Predictor (OBP) of θ_i. Since the BPEs do not appeal to the assumed linking model, the associated OBPs may be more robust to misspecification of the linking model than the customary EBs. Empirical results showed that under correct specification of the linking model, the OBP and EB estimators perform similarly, and lead to considerable efficiency gains when the linking model is not correctly specified.

Estimation of MSE of OBP estimator of θ_i is problematic because the assumed linking model is misspecified. A way around this difficulty is to estimate the conditional MSE of the OBP given θ, similar to (3.7) for the EB estimator. Jiang et al. (2011) proposed a second-order unbiased estimator of the conditional

MSE of OBP but it involves the term $(\hat{\theta}_i^{OBP} - \hat{\theta}_i)^2$ similar to the term $(\hat{\theta}_i^{EB} - z_i^T \hat{\beta})^2$ in (3.7). As a result, the proposed MSE estimator can be highly unstable as in the case of (3.7).

Benchmarking methods. It is desirable in practice to ensure that the model-based estimators of area means when aggregated agree with a reliable direct estimator. If θ_i is the area mean, then the EB estimators $\widehat{\theta}_i^{EB}$ of area means do not satisfy this benchmarking property in the sense $\sum_{t=1}^{m} W_t \hat{\theta}_t^{EB} \neq \sum_{t=1}^{m} W_t \hat{\theta}_t, = \hat{\theta}_+$, where W_t is the known proportion of units in area t and $\hat{\theta}_+$ is the direct estimator of the aggregate mean.

Simple adjustments to the EB estimators to satisfy benchmarking include ratio benchmarking and difference benchmarking respectively given by

$$\hat{\theta}_i^{RB} = \widehat{\theta}_i^{EB} (\hat{\theta}_+ / \sum W_t \hat{\theta}_t^{EB}) \tag{3.9}$$

and

$$\hat{\theta}_i^{DB} = \hat{\theta}_i^{EB} + (\hat{\theta}_+ - \sum W_t \hat{\theta}_t^{EB}). \tag{3.10}$$

Steorts and Ghosh (2013) derived a second-order unbiased estimator of $MSE(\hat{\theta}_i^{DB})$ given by $\text{mse}(\hat{\theta}_i^{DB}) = mse(\hat{\theta}_i^{EB}) + g_4(\hat{\sigma}_v^2)$, where the common term $g_4(\hat{\sigma}_v^2)$ is positive. This result shows that the effect of benchmarking is to increase the MSE. However, in their application to estimation of poor school age children in the USA they found negligible inflation in MSE due to difference in benchmarking.

A limitation of RB and DB estimators is that a common adjustment factor is applied to all the EB estimators regardless of their precision. Alternative benchmarked estimators that avoid the above limitation have been proposed (Wang, Fuller and Qu (2008) and Datta et al. (2011b). Bell, Datta and Ghosh (2013) extended the Wang et al. method to multiple benchmark constraints. Two alternative methods (Wang, Fuller and Qu 2008) and You, Rao and Hidiroglou 2013) provide self-benchmarking estimators of area means in the sense that estimators that automatically satisfy the benchmarking constraint are obtained. The method of You et al. (2013) replaces the estimator of β used in the EB estimator by an alternative estimator that depends on the benchmarking weights W_t. On the other hand, the method of Wang et al. (2008) replaces the covariate vector z_i^T by $(z_i, W_i \psi_i)^T$ in the linking model (3.2) and then uses the EB estimator of the area mean based on the augmented model. An advantage of both methods is that MSE estimation requires no new theory.

4. Basic unit level nested error models

4.1. Estimation and MSE estimation

In some applications, for example business surveys, unit level sample data $\{(y_{ij}, x_{ij}), j = 1,..., \mathrm{n}_i; i = 1,..., m\}$ are often available for the sampled areas, where n_i is the sample size in area i. We assume that the area population means $\bar{X}_i$ of the auxiliary variables x_{ij} are known for the estimation of area means $\bar{Y}_i$.

For the estimation of complex non-linear parameters, such as poverty measures, we need to know all the population values $x_{ij}, j = 1,..., N_i$, where N_i is the number of population units in area i. We assume a basic unit level nested error model for the population and assume that the same model holds for the sample (Battese, Harter and Fuller 1988):

$$y_{ij} = x_{ij}^T \beta + v_i + e_{ij}, \tag{4.1}$$

where $v_i \sim_{\text{iid}} N(0, \sigma_v^2)$ are random area effects independent of unit errors $e_{ij} \sim_{\text{iid}} N(0, \sigma_e^2)$. Under the above set-up, unit level models can lead to significant efficiency gains over area level models, because the model parameters $(\beta, \sigma_v^2, \sigma_e^2)$ can be estimated more accurately using all the $n = \sum n_i$ unit level observations. In some applications, it is more realistic to assume unequal error variances $\sigma_{eij}^2 = k_{ij}^2 \sigma_e^2$, where k_{ij} is a known constant (Stukel and Rao 1999). For example, in business surveys with a scalar covariate x_{ij}, the choice $k_{ij}^2 = x_{ij}$ is often used.

The area mean $\bar{Y}_i$ may be approximated by $\mu_i = \bar{X}_i^T \beta + v_i$, assuming that N_i is large. Then, the best estimator of μ_i is given by

$$\tilde{\mu}_i^B = \gamma_i [\bar{y}_i + (\bar{X}_i - \bar{x}_i)^T \beta] + (1 - \gamma_i)(\bar{X}_i^T \beta), \tag{4.2}$$

where $(\bar{y}_i, \bar{x}_i)$ are the area sample means and $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2 / n_i)$. The estimator (4.2) is a weighted combination of the sample regression estimator of $\bar{y}_i + (\bar{X}_i - \bar{x}_i)^T \beta$ and the regression synthetic estimator $\bar{X}_i^T \beta$. In practice, we replace the model parameters by suitable estimators $(\hat{\beta}, \hat{\sigma}_v^2, \hat{\sigma}_e^2)$, in particular REML estimators and the resulting EB estimator is denoted by $\hat{\mu}_i^{EB}$.

Note that (4.2) does not take account of survey weights, w_{ij}, unlike the EB estimator (3.2) under the area level model. As a result, it is not design consistent as the area sample size increases, unless the weights are equal within each area. It is desirable to ensure design consistency because n_i could be moderately large for some of the areas, for example California when the US states are regarded as areas. A pseudo-EB estimator, proposed by You and Rao (2002), avoids this difficulty by taking account of weights and at the same time ensuring self-benchmarking.

Estimation of $\mathrm{MSE}(\hat{\mu}_i^{EB})$ has received considerable attention, and second order unbiased MSE estimators have been derived using Taylor linearization, jackknife and bootstrap methods. Hall and Maiti (2006) relaxed the normality assumption of model (4.1) and obtained second order unbiased MSE estimators using a double-bootstrap method that matches the estimated second and fourth moments of v_i and e_{ij}. The first phase bootstrap samples are used to obtain a first order MSE estimator, similar to (3.6) for the area level model, and its bias is then corrected using the second phase bootstrap samples. Regarding the choice of first phase and second phase bootstrap sample sizes, B_1 and B_2, Fuller and Erciulescu (2014) demonstrated that the choice $B_2 = 1$ and $B_1 = R/2$ leads to smaller bootstrap error than other choices of B_1 and B_2, where $R = B_1(B_2 + 1)$ is the total number of bootstrap replicates. This result implies that one should select a single second phase bootstrap sample from each first phase bootstrap sample. Pfeffermann and Correa (2012) studied efficient methods of bootstrap MSE estimation for the normal case and proposed an empirical bootstrap bias correction method that performed significantly better than the Hall-Maiti method.

4.2. Practical issues

As in the case of the FH model, we need to address practical issues in implementing EB estimation under the basic unit level model (4.1). Those issues include (i) model misspecification, (ii) robust estimation in the presence of outliers, (iii) estimation of complex parameters, (iv) measurement errors in the covariates and (v) informative sampling. We give a very brief account of methods proposed to deal with the above issues.

Model misspecification: Jiang, Nguyen and J. S. Rao (2014) extended their OBP method to the nested error model and studied its performance under misspecification of either the mean function $m(x) = x^T\beta$ or the variance of the unit error e_{ij} or both , assuming simple random sampling within areas. They also proposed a bootstrap estimator of MSE of the OBP estimator of area mean. An alternative approach to dealing with misspecification of mean function is to use a semi-parametric nested error model with unspecified mean function $m(x)$. Opsomer et al. (2008) used a truncated polynomial spline basis to approximate the mean function for the scalar x case and showed that it leads to a linear mixed model but it does not have a block diagonal covariance structure unlike model

(4.1). They obtained the EB estimators of area means and also proposed a bootstrap estimator of MSE.

Robust estimation: Estimation of area means that are robust to outliers in the random effects v_i and/or unit errors e_{ij} has received considerable attention in recent years. Sinha and Rao (2009) proposed robust EBLUP estimators and associated bootstrap MSE estimators. Their results suggest that the customary EBLUP (or EB) is robust to outliers in v_i but not to outliers in e_{ij}. They assumed mean zero random effects and unit errors. Computational issues associated with the Sinha-Rao method are addressed in Schoch (2012). Rao, Sinha and Dumitrescu (2014) extended robust EBLUP estimation to the semi-parametric spline models. Chambers et al. (2014) studied bias-adjusted robust estimators and associated MSE estimators using area-specific residuals. Jiango, Haziza and Duchesne (2014) developed efficient bias corrections using all the sample residuals.

An alternative approach to REBLUP is the M-quantile method (Chambers and Tzavidis 2006). The method uses unit level data and assumes that all "M-quantiles" of the conditional distribution of y given x are linear in x, but random area effects are not directly incorporated into the model. Tzavidis and Chambers (2005) studied bias-adjusted M-quantile estimators.

Estimation of complex area parameters. Estimation of complex parameters, in particular poverty measures (poverty rate, poverty gap and poverty severity) has received considerable attention in recent years because of growing demand for reliable area-level poverty indicators. Molina and Rao (2010) developed EB estimators for complex parameters under a nested error model that uses log (welfare variable) as y. The EB method performed significantly better than a "simulated census" method widely used by the World Bank (WB) for poverty mapping in developing countries. Diallo and Rao (2014) relaxed the normality assumption by using skew normal (SN) distributions on v_i and/or e_{ij}. Their results indicate that the normality based EB estimators are sensitive to non-normality of e_{ij} but not to non-normality of v_i. Berg and Chandra (2014) also used nested error models for the log of the variable of interest, but their focus was on estimating area means of the variable of of interest.

Measurement errors in covariates. Ghosh and Sinha (2007) formulated a functional measurement unit level error model with a scalar area level covariate x_i subject to measurement errors. They assumed that independent values x_{ij} of the true x_i are measured such that x_{ij} corresponds to y_{ij}. Under this set-up they obtained a pseudo-EB estimators of area means. Datta, Rao and Torabi (2010) obtained more efficient pseudo-EB estimators by making fuller use of the available data. A more realistic model assumes that the x_{ij} values are drawn from an independent survey (Arima, Datta and Liseo 2014). Ghosh, Sinha and Kim (2006) and Torabi, Datta and Rao (2009) studied structural measurement error models with stochastic x_i.

Informative sampling. Most of the recent SAE papers assumed non-informative sampling in the sense that the assumed population model also holds for the sample. Under informative sampling, the survey design is related to the variable of interest given the predictor variables in the model, and in this case population model may not hold for the sample data. The pseudo-EB estimator of Rao and You (2012) uses the survey weights to ensure design consistency, but it is derived under non-informative sampling. However, empirical results suggest that it performs quite well in terms of bias under informative sampling unlike the EB estimator that ignores survey weights (Stefan 2005, Verret, Rao and Hidiroglou 2015).

Pfeffermann and Sverchkov (2007) proposed a bias-adjusted EB estimator for unit level models under informative sampling by modelling the conditional expectation of sampling weights given the sample as a function of y and x. They also studied the case of informative sampling of areas and units within areas. An alternative approach, when all areas are sampled, augments the unit level model (4.1) by including a suitable function of the selection probability p_{ij} of unit (ij) as an additional covariate g_{ij} and then uses standard EB estimators based on the augmented model (Verret, Rao and Hidiroglou 2015). The augmented model approach performed well in empirical studies, but it assumes that the population mean, $\bar{G}_i$, of the augmented variable is known. The selection of the augmenting variable may be based on plots of model (4.1) residuals against different choices of g_{ij}. In particular, if $g_{ij} = p_{ij}$ is a suitable choice, then the mean $\bar{G}_i = N_i^{-1}$ is known.

5. Model selection and checking

Model-based small area estimation heavily depends on the validity of the assumed model for the sample data. It is therefore important to use appropriate methods for model selection and then do checking of the selected model through residual analysis, influential diagnostics, etc. Most of the recent literature on model selection assumes non-informative sampling. Variable selection is an important component of model selection. Recent methods for variable selection in linear mixed models include fence methods (Jiang, J. S. Rao, Gu and Nguyen 2008), conditional AIC for predictive performance (Vaida and Blanchard 2005) and Han (2011) for the FH model. Muller, Scealy and Welsh (2013) present a comprehensive review of model selection in linear mixed models. One major problem with existing model diagnostics is the assumption of non-informative sampling. If sampling is informative, then the identified sample model may not hold for the population and hence it can lead to erroneous inferences. The augmented model approach of Verret et al. (2015) might be a way to get around this difficulty because the identified sample augmented model also holds for the population. Alternatively, the approach of Pfeffermann and Sverchkov (2007) to deal with informative sampling only requires fitting the model holding for the sample data and the sample model for the weights. Hence, the previous model diagnostics should apply under their approach. Pfeffermann (2013) reviewed recent method for model selection and checking. Both internal evaluations through model diagnostics and external evaluations, based on comparing estimates derived from models with reliable values obtained from external sources, play an important role in small area estimation.

6. Concluding remarks

We have focused on recent important developments related to the basic area level and unit level models and highlighted some practical issues in implementing model-based small area estimation, in particular EB (or EBLUP) methods. Due to space limitations, hierarchical Bayes (HB) method, based on assumed priors on model parameters, is not covered in this paper. The longest chapter in the author's 2003 book is on the HB approach to SAE. It is a powerful approach and provides "exact" inferences for complex models. Also, we did not include recent developments in SAE based on generalized linear mixed models (GLMMs) used for unit level binary or count data. Many recent extensions of the basic models are also not covered in this paper. SAE is experiencing explosive growth and we will see many important new developments in both theory and applications in the next 10 years. Review papers on SAE in the past 10 years include Rao (2005, 2008), Jiang and Lahiri (2006), Datta (2009) and Pfeffermann (2013).

REFERENCES

ARIMA, S., DATTA, G. S., LISEO, B., (2015). Accounting for measurement errors in SAE: an overview, in *Analysis of Poverty Data by Small Area Methods,* M. Pratesi (Ed.), Hoboken: Wiley (in press).

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, 28−36.

BELL, W. R., DATTA, G. S., GHOSH, M., (2011). Benchmarking small area estimators. Biometrika, 100, 189−202.

BERG, E., CHANDRA, H., (2014). Small area prediction for a unit-level lognormal model. *Computational Statistics and Data Analysis*, 78, 158-175.

CASAS-CORDERO, C., ENCINA, J., LAHIRI, P.,(2015). Poverty mapping for the Chilean Comunas, in *Analysis of Poverty data by Small Area Methods,* M. Pratesi (Ed.), Hoboken: Wiley (in press).

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

CHAMBERS, R., TZAVIDIS, N., (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255−268.

CHATTERJEE, S., LAHIRI, P., LI, H., (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Annals of Statistics,* 36, 1221−1245.

DATTA, G. S., (2009). Model-based approach to small area estimation, in *Sample Surveys: Inference and Analysis,* D. Pfeffermann and C. R. Rao (Eds.), Vol. 29B, Amsterdam: North-Holland, pp. 251−288.

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 measurement errors. *Journal of Statistical Planning and Inference,* 140, 2952−2962.

DATTA, G. S., KUBOKAWA, T., MOLINA, I., RAO, J. N. K., (2011a). Estimation of mean squared error of model-based small area estimators. *Test,* 20, 367−388.

DATTA, G. S., GHOSH, M., STEORTS, S., MAPLES, J. J., (2011b). Bayesian benchmarking with applications to small area estimation. *Test,* 20, 574−588.

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 Methods Section,* American Statistical Association.

ERCIULESCU, A. L., FULLER, W. A.,(2014). Parametric bootstrap procedures for small area prediction variance. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

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.

GHOSH, M., SINHA, K.,(2007). Empirical Bayes estimation in finite population sampling under functional measurement error models. *Scandinavian Journal of Statistics,* 33, 591−608.

GHOSH, M., SINHA, K., KIM, D.,(2006). Empirical and Hierarchical Bayesian estimation in finite population sampling under structural measurement error models. *Journal of Statistical Planning and Inference,* 137, 2759−2773.

HALL, P., MAITI, T., (2006). Nonparametric estimation of mean-squared prediction error in nested error regression models. *Annals of Statistics,* 34, 1733−1750.

HAN, B., (2013). Conditional Akaike information criterion in the Fay-Herriot model. *Statistical Methodology,* 11, 53−67.

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

JIANG, J., RAO, J. S., GU, I., NGUYEN, T., (2008). Fence Methods for Mixed Model Selection. *Annals of Statistics*, 36, 1669−1692.

JIANG, J., RAO, J. S., NGUYEN, T., (2011). Best predictive small area estimation. *Journal of the American Statistical Association,* 106, 732−745.

JIANG, J., NGUYEN, T., RAO, J. S., (2014). Observed best prediction via nested –error regression with potentially misspecified mean and variance. *Survey Methodology (in press).*

JIANGO, V. D., HAZIZA, D., DUCHESNE, P.,(2013). Controlling the bias of robust small-area estimation. *Biometrika*, 100, 843−858.

LI, H., LAHIRI, P., (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis,* 101, 882−892.

LUERY, D. M., (2011). Small area income and poverty estimates program. *Proceedings of 27th SCORUS Conference,* Jurmala, Latvia, pp. 93−107.

MARCHETTI, S., GIUSTI, C., PRATESI, M., SALVATI, N., GIANNOTTI, F., PEDRESCHI, D., RINIZIVILLO, S., PAPPALARDO, L., GABRIELLI, L., (2015). Small area model based estimation using big data sources. *Journal of Official Statistics,* to appear.

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

MOLINA, I., RAO, J. N. K., DATTA, G. S., (2014). Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random effects. *Survey Methodology* (in press).

MULLER, S., SCEALY, J. L., WELSH, A. H., (2013). Model selection in linear mixed models. *Statistical Science,* 28, 135−167.

OPSOMER, J. D., CLAESKENS, G., RANDALL, M.G., KAUERMANN, G., BREIDT, F. J.,(2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Ser. B,* 70, 265−286.

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

PFEFFERMANN, D., SVERCHKOV, M.,(2007). Small-area estimation under informative probability sampling of areas and within selected areas. *Journal of the American Statistical Association,* 102, 1427−1439.

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

PRATESI , M., (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics* (in press).

RANDRIANASOLO, T., TILLE, Y.,(2013). Small area estimation by splitting the sampling weights. *Electronic Journal of Statistics,* 7, 1835−1855.

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

RAO, J. N. K., (2005). Inferential issues in small area estimation: some new developments. *Statistics in Transition,* 7, 513−526.

RAO, J. N. K., (2008). Some methods for small area estimation. *Rivista Internazionale di Scienze Sociali,* 4, 387−406.

RAO, J. N. K., SINHA, S. K., DUMITRESCU, L., (2014). Robust small area estimation under semi-parametric mixed models. *Canadian Journal of Statistics,* 42, 126−141.

RIVEST, L-P., VANDAL, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling.* Technical Report No. 386, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, Canada.

SCHIRM, A. L., ZASLAVSKY, A. M., (1997). Reweighting households to develop micro simulation estimates for states. *Proceedings of the 1997 Section on Survey Research Methods,* American Statistical Association, pp. 306−311.

SCHOCH, T., (2012). Robust unit-level small area estimation: a fast algorithm for large data sets. *Austrian Journal of Statistics,* 41, 243−265.

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

STEORTS, R., GHOSH, M., (2013). On estimation of mean squared errors of benchmarked empirical Bayes estimators. *Statistica Sinica,* 23, 749−767.

STUKEL, D. M., RAO, J. N .K.,(1999). Small-area estimation under two-stage nested error regression models. *Journal of Statistical Planning and Inference,* 78, 131−147.

TORABI, M., DATTA, G. S., 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.

TZAVIDIS, N., CHAMBERS, R., (2005). Bias adjusted small area estimation with M-quantile models. *Statistics in Transition,* 7, 707−713.

VAIDA, F., BLANCHARD, S., (2005). Conditional Akaike information for mixed effect models. *Biometrika,* 92, 351−370.

VERRET, F., RAO, J. N. K., HIDIROGLOU, M. A., (2014). Model-based small area estimation under informative sampling. *Survey Methodology* (in press).

WANG, J., FULLER, W. A., (2003). The mean squared error of small area predictors constructed with estimated sampling variances. *Journal of the American Statistical Association,* 98, 718−723.

WANG, J., FULLER, W. A., QU, Y., (2008). Small area estimation under a restriction. *Survey Methodology,* 34, 29−36.

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

YOSHIMORI, M., LAHIRI, P., (2014). A new adjusted maximum likelihood method for the Fay-Herriot small area model. *Journal of Multivariate Analysis*, 124, 281−294.

YOU, Y., RAO, J. N. K., (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation under survey weights. *Canadian Journal of Statistics*, 30, 431−439.

YOU, Y., RAO, J. N. K., HIDIROGLOU, M. A., (2013). On the performance of self-benchmarked small area estimates under the Fay-Herriot area level model. *Survey Methodology*, 39, 217−229.

STATISTICS IN TRANSITION new series and SURVEY METHODOLOGY
Joint Special Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 511–522

TRIPLE-GOAL ESTIMATION OF UNEMPLOYMENT RATES FOR U.S. STATES USING THE U.S. CURRENT POPULATION SURVEY DATA

Daniel Bonnéry [1] Yang Cheng[2] Neung Soo Ha[3] Partha Lahiri[4]

ABSTRACT

In this paper, we first develop a triple-goal small area estimation methodology for simultaneous estimation of unemployment rates for U.S. states using the Current Population Survey (CPS) data and a two-level random sampling variance normal model. The main goal of this paper is to illustrate the utility of the triple-goal methodology in generating a single series of unemployment rate estimates for three separate purposes: developing estimates for individual small area means, producing empirical distribution function (EDF) of true small area means, and the ranking of the small areas by true small area means. We achieve our goal using a Monte Carlo simulation experiment and a real data analysis.

Key words: complex survey data; empirical distribution function; Monte Carlo Markov Chain; rank; risk; small area estimation,

1. Introduction

The national unemployment rate is one of the five key economic indicators published by the United States Bureau of Labor Statistics (BLS) and represents the number of unemployed as a percentage of the labor force. BLS publishes unemployment rate estimates for the nation and its different demographic and geographic subdomains. For example, unemployment rate estimates are produced for all states and the District of Columbia, all metropolitan statistical areas (MSA), all counties, cities and towns of New England, and all cities with population 25,000 or greater. The local unemployment rate estimates are used for regional planning and fund allocation under various federal assistance programs. The primary source of data for the unemployment rate statistics for both small and large domains is the Current Population Survey (CPS) conducted by the Census Bureau for BLS. The data are collected for about 729 MSAs consisting of more than 1,000 counties covering every state and the District of Columbia. More information about the CPS can be found at `http://www.bls.gov/cps/`

[1] Joint Program in Survey Methodology, University of Maryland. E-mail: dbonnery@umd.edu

[2] U.S. Census Bureau. E-mail: yang.cheng@census.gov

[3] Nielsen. E-mail: Neung.Ha@nielsen.com

[4] Joint Program in Survey Methodology, University of Maryland. E-mail: plahiri@umd.edu

The Census Bureau has been using the so-called AK composite estimation technique for generating national employment and unemployment levels and rates for the last several decades. The AK composite estimation technique, developed using the ideas of Gurney and Daly (1965), essentially improves on the standard survey-weighted estimates by borrowing strength over time. For more information on the AK estimation, see Lent et al. (1999). The estimation methodology for the BLS Local Area Unemployment Statistics (LAUS) can be found at http://www.bls.gov/lau/laumthd.htm. The state level unemployment statistics are based on a paper by Pfeffermann and Tiller (2006). For each month, model-based census division estimates are first benchmarked to the non-seasonally adjusted national A-K composite estimate and then similar model-based state estimates are benchmarked to the benchmarked estimate of the state's division. The unemployment estimates for the states or the census divisions can be viewed as benchmarked empirical best prediction (EBP) estimates, derived using a state-space model and implemented via an innovative Kalman Filter updating scheme that simplifies the computational burden in a complex production environment.

In a statistical decision-theoretic framework, BLS addresses the problem of point estimation under a squared error loss function and the estimation of the corresponding risk measured by the mean squared prediction error (MSPE). These are indeed important statistical decision problems. It is expected that BLS will continue to focus on the point estimation and the corresponding MSPE because of a long history of such unemployment statistics series and official publication requirements. One can, however, envision a variety of statistical decision problems related to unemployment statistics. For example, different stakeholders may be interested in ranking different states in order of unemployment rates or identifying states with unemployment rates exceeding a certain specified threshold for regional planning and fund allocation problems. The need for answering research questions other than point estimation can be found in different contexts. For example, the goal can be estimating the performance evaluation, like the rank, among different companies; see Landrum et al. (2000). Reporting an ensemble of estimates can also provide useful interpretation in disease mapping to ascertain variation in disease rates for different geographical regions; see Conlon and Louis (1999) and Devine and Louis (1994).

Note that the research questions mentioned above correspond to different s-tatistical decision-theoretic problems and thus, statistically speaking, a research question-specific unemployment series can be found, which is likely to be different from BLS published series. Of course, the published unemployment rates can be used to answer a variety of research questions, but they may not be well suited for a wide range of problems. To elaborate this point, if the ranks of parameters are the target, under the Bayesian approach, the conditional expected ranks are optimal under squared error loss function, but ranking posterior means, which are optimal for point estimation under squared error loss, can perform poorly; see Goldstein and Spiegelhalter (1996). If the feature of interest is the histogram or the empirical distribution function (EDF) of the parameters, then the conditional expected EDF is optimal under integrated squared error loss function, and the histogram of the posterior means of the parameters is underdispersed; see Ghosh (1992). There are a number of papers on the estimation of parameters for an individual small area, e.g. Rao (2003), Jiang and Lahiri (2006), Pfeffermann (2013), a histogram of small area parameters, e.g. Louis (1984), Lahiri (1990), Ghosh (1992), and ranking small area parameters, e.g. Laird and Louis (1989).

Although different series can be produced to address different questions, reporting several ensembles for all different situations would be inefficient and may cause inconsistencies. While there does not exist a set of point estimates that simultaneously optimize all of these criteria (Gelman and Price, 1999), Shen and Louis (1998) developed an interesting method, called "triple-goal" estimation method, which produces estimates that perform reasonably well with respect to all three criteria.

In Section 2, we explore a triple-goal small area estimation methodology for simultaneous estimation of small area means using the CPS complex survey data. The main goal is to produce a set of small area estimates that are good for simultaneously meeting three different goals of developing estimates for individual small area means, producing histogram of true small area means, and ranking of the s-mall areas by true small area means. We discuss evaluation of our methodology in Section 3.

2. Adaptation of the triple-goal estimation methodology to estimate unemployment rates for U.S. states

The main challenge for adapting the existing triple-goal methodology to estimate unemployment rates for U.S. states is to incorporate the complex survey features of the CPS. Let $\hat{\pi}_i$ be the survey-weighted direct estimate of the true unemployment rate π_i for the ith state $(i = 1, \cdots, m)$. We are interested in producing triple-goal estimates of $\pi = (\pi_1, \cdots, \pi_m)$. To obtain triple-goal estimates of π_i's and to compare with the corresponding Bayesian estimates (posterior means of π_i's), we consider the following hierarchical model.

For $i = 1, \ldots, m$,

$$\textit{Level 1} \text{ (sampling distribution)} : \hat{\pi}_i | \pi_i \overset{ind}{\sim} N\left(\pi_i,\ \frac{\pi_i(1-\pi_i)}{n_{i;\text{eff}}}\right);$$

$$\textit{Level 2} \text{ (prior distribution)} : \text{logit}(\pi_i) | \mu, A \overset{iid}{\sim} N(\mu,\ A),$$

where π_i and $n_{i;eff}$ are the "true" unemployment rate and the effective sample size for state i, respectively. The effective sample size for a state is the ratio of the CPS sample size for that state and the national estimate of design effect (deff). We assume flat priors on both μ and A.

We note that the BLS uses a two-level time series normality-based model to combine previous survey data. While the BLS model will be of interest to produce triple-goal unemployment rate estimates, in this paper we focus on the above relatively simple cross-sectional random sampling variance two-level normal model for demonstrating the utility of triple-goal estimation for multi-purpose estimation. Like the BLS model, we find it convenient to assume normality for the survey-weighted proportions, but use a random sampling model to incorporate uncertainty in estimating sampling variances of the survey weighted proportions. Such a model was considered earlier in different contexts by Liu et al. (2014) and Ha et al. (2014).

The triple-goal estimation method involves the following three steps (see Shen and Louis (1998) for further details):

Step 1: Produce element-specific point estimates with "optimality" qualities for the region of interest;

Step 2: Obtain an ensemble of point estimates that best approximate the histogram of the true parameter ensemble; see Louis (1984);

Step 3: Rank within a selected ensemble.

The procedure for obtaining triple-goal estimators follows along the line of Shen and Louis (1998), which is described below:

First, we need to obtain an estimate of the empirical distribution function (EDF) of π. The EDF of π is defined as:

$$F_m(\alpha) = \frac{1}{m} \sum_{j=1}^{m} \mathscr{I}\{\pi_j \leq \alpha\},$$

where $\alpha \in \mathbb{R}$ and $\mathscr{I}$ is the indicator function. Under the following integrated squared error loss (ISEL) function for a given EDF estimator $\tilde{F}_m$:

$$\text{ISEL}(F_m, \tilde{F}_m) = \int \left[F_m(\alpha) - \tilde{F}_m(\alpha)\right]^2 d\alpha,$$

the Bayes estimator of EDF is given by

$$\hat{F}_m(\alpha) = E\left[F_m(\alpha)|\hat{\pi}\right] = \frac{1}{m}\sum_{j=1}^{m} P(\pi_j \leq \alpha|\hat{\pi}).$$

Secondly, we need to obtain the rank of the parameter ensemble π. The rank of π_i is defined as

$$R_i = \text{rank}(\pi_i) = \sum_{j=1}^{m} \mathscr{I}\{\pi_i \geq \pi_j\}.$$

Under the rank squared error loss (RSEL) function for a given rank estimator $\tilde{\mathbf{R}}$, defined as

$$\text{RSEL}(\mathbf{R}, \tilde{\mathbf{R}}) = \frac{1}{m}\sum_{j=1}^{m} (R_j - \tilde{R}_j)^2,$$

the Bayes estimator of R_i is given by

$$\bar{R}_i = E(R_i|\hat{\pi}) = \sum_{j=1}^{m} P(\pi_i \geq \pi_j|\hat{\pi}).$$

The $\bar{R}_i$'s are not integers in general; however, it is easy to transform them in order and denote it by:

$$\hat{R}_i = \text{rank}(\bar{R}_i|\mathbf{R}), i = \ldots, m.$$

Finally, we generate an ensemble of point estimates, conditional on the optimal estimate of the ensemble EDF, $\hat{F}_m$, and the optimal estimates of the ranks, $\hat{R}_i$. Furthermore, the added constraint that $\hat{F}_m$ is a discrete distribution with at most m mass points, the triple-goal estimator is defined as:

$$\hat{\pi}_i^{TG} = \hat{F}_m^{-1}\left(\frac{2\hat{R}_i - 1}{2m}\right), i = 1, \ldots, m.$$

We use MCMC to implement the triple-goal method. The simulated samples after deleting the first B “burn-in" samples, i.e.

$$\left\{\mu^{(B+\ell)}, A^{(B+\ell)}, \pi^{(B+\ell)},\ \ell = 1, \cdots, L\right\},$$

are considered as L simulated samples from the posterior distribution of β, A, π.

The posterior density of π is approximated by

$$\left\{\pi^{(B+\ell)},\ \ell = 1, \cdots, L\right\}.$$

In particular, we need the following approximations:

$$\hat{F}_m(\alpha) \approx \frac{1}{m}\sum_{j=1}^{m}\left\{\frac{1}{L}\sum_{\ell=1}^{L}\mathscr{I}\left[\pi_j^{(B+\ell)} \leq \alpha\right]\right\},$$

$$\bar{R}_i \approx \sum_{j=1}^{m}\left\{\frac{1}{L}\sum_{\ell=1}^{L}\mathscr{I}\left[\pi_i^{(B+\ell)} \leq \pi_j^{(B+\ell)}\right]\right\}.$$

3. Evaluation

Our ultimate goal is to develop a triple-goal estimation system for the state unemployment rates using the CPS data. As in any real life data analysis, we encounter the challenging problem of evaluation of triple-goal estimates relative to the commonly used direct and posterior means since we do not have true unemployment values. We consider two options. First, we compare different estimates using simulated data generated using the model given in Section 2 and the CPS data. While such an evaluation is model-dependent, we argue that this is a reasonable approach since our main goal in this paper is to compare direct estimates, posterior means and triple-goal estimates for three separate purposes given a working model. In Subsection 3.1, we present results from such an evaluation study. The other option for evaluation is to use a real data that contain the truth or a gold standard. We do not have such data for unemployment rate estimation research. Since estimation of unemployment rates is essentially a problem of estimation of proportions, in Subsection 3.2 we use the well-known batting average data described in Efron and Morris (1975), which contain true batting averages (true proportions).

We now evaluate direct, posterior mean, triple-goal estimators of ranks, EDFs, and individual parameters. To be specific, we compare different estimators using the following four summary evaluation measures:

(i) Root Average Squared Deviation (RASD): $\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\tilde{\pi}_i - \pi_i)^2}$

(ii) Root Integrated Squared Error Loss (RISEL): $\sqrt{\int\left[F_m(t) - \tilde{F}_m(t)\right]^2 dt}$

(iii) Variance Ratio (VR): $\frac{\sum_{i=1}^{m}(\tilde{\pi}_i - \bar{\tilde{\pi}})^2}{\sum_{i=1}^{m}(\pi_i - \bar{\pi})^2}$

(iv) Root Rank Average Squared Deviation (RRASD): $\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\tilde{R}_i - R_i)^2}$,

where $\bar{\pi}_i$ ($\bar{\tilde{\pi}}$) is the average of the π_i's ($\tilde{\pi}_i$'s), average being taken over all m states.

3.1. Evaluation using simulated data

Using the two-level normal model described earlier with $\mu = \hat{\pi}$, the national unemployment rate estimate, and $A = \sum_{i=1}^{51}(\hat{\pi}_i - \bar{\hat{\pi}})^2/51$, where $\hat{\pi}_i$ is the survey-weighted CPS unemployment rate for state i $(i = 1, \cdots, m)$, we generate unemployment rate direct estimates and simulated true values for the states. We can then compare different methods using simulated values.

Table 1 displays values of the four evaluation measures for the three estimators. From the VR measure, it is clear that the variability of the direct estimates of the state unemployment rates overestimates the corresponding variability of the simulated unemployment rates across the states. On the other hand, the posterior means of the state unemployment rate estimates overshrink. The triple-goal estimates are almost perfect in terms of this criterion. Based on the RISEL criterion, the triple goal estimates are also the best among the three sets of estimates in terms of estimating the EDF of the simulated unemployment rates. The criterion RRASD suggests that in terms of the rank, triple-goal estimates are the best, but they are only marginally better than the posterior means. In terms of the RASD criterion, posterior means are the best as expected, but are only marginally better than the triple-goal estimates.

Figure 1 provides histograms of three sets of estimates for the states and the simulated values. From a visual inspection, it is clear that the histogram for the triple-goal estimates is the closest to that of the true values when compared to the histograms of the posterior means and the direct estimates.

	RASD	RISEL	VR	RRASD
direct	0.0097	0.0122	1.2861	8.2652
post. mean	0.0086	0.0121	0.8316	8.1889
triple-goal	0.0095	0.0091	1.0200	8.1746

Table 1: Summary statistics for the unemployment data

3.2. Evaluation using a real data with true values

As mentioned before, in this subsection we use the well-known baseball data, which were used earlier by researchers in evaluating different small area methodologies. The data contain batting averages of eighteen major league baseball players in the 1970 season. Each player had batted 45 times and their batting averages are recorded up to that point. Using this data alone, Efron and Morris (1975) wanted to predict each player's batting average for the remainder of the 1970 season. Here, a player corresponds to a small area like a state in the unemployment rate estimation.

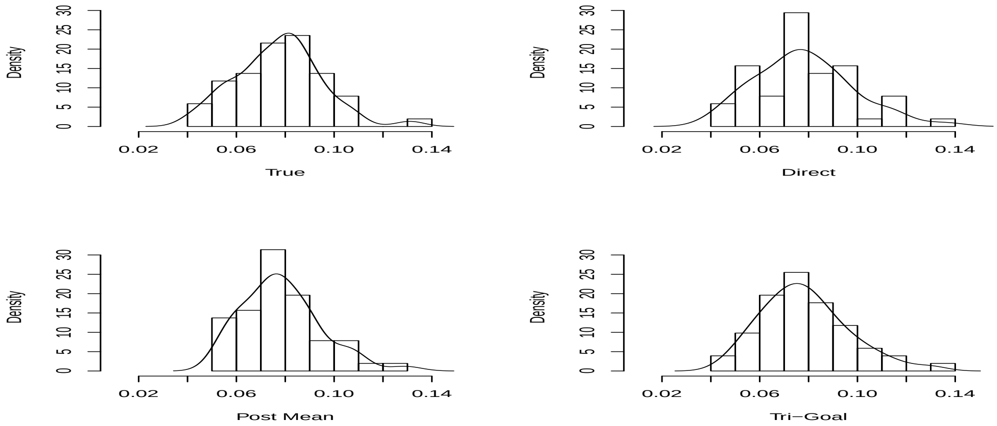

Figure 1: Histograms for the unemployment data

We report the four summary evaluation measures for the three sets of estimates in Table 2. In Figure 2, we plot the histograms for the three sets of estimates of batting averages and the true batting averages. The conclusion is similar to the one in Subsection 3.1.

	RASD	RISEL	VR	RRASD
direct	0.0572	0.0486	3.3920	5.8878
post.mean	0.0311	0.0311	0.1899	5.8214
triple-goal	0.0334	0.0094	1.0328	5.8022

Table 2: Summary statistics for the baseball data

4. Concluding Remarks

In this paper, we extend the triple-goal methodology, originally proposed by Shen and Louis (1998), to a hierarchical model not considered earlier in modeling unemployment rates for small areas. First, instead of using fixed and known sampling variance of a survey-weighted unemployment rate for a small area, we have used the true variance formula of a sample proportion with sample size replaced by the effective sample size in order to incorporate the complex survey design. Secondly, to borrow strength from small areas, we use normality on the logistic function of the unknown true unemployment rates, which appear in both the means and variances in the sampling distribution.

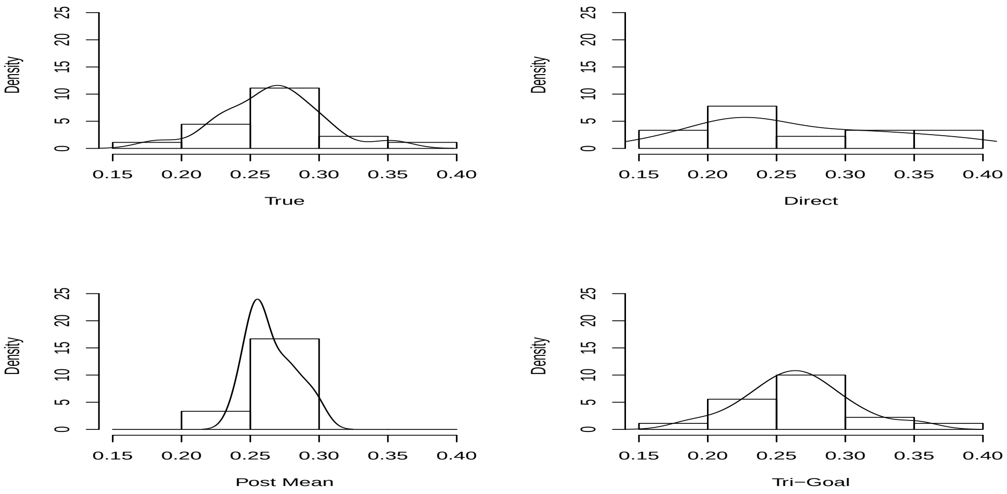

Figure 2: Histograms for the baseball data

We reiterate that the triple-goal method is for multi-purpose inferences. In theory, this approach should reduce the overshrinking problem associated with the standard Bayesian estimates (posterior means) targeted for point estimation and should do better than rival methods in estimating ranks and empirical distribution function of the true values. While our evaluation studies demonstrate a clear superiority of the triple-goal method in reducing the overshrinking problem and estimating the empirical distribution function of the true values, it is only marginally better than the posterior means and direct estimates in estimating ranks. This could be due to certain approximations applied to the optimal rank estimates in order to produce integer valued ranks of the small areas. Under the theoretical setting, posterior means should perform better than the triple-goal estimates in terms of point estimation of the small area proportions. Our evaluation studies, however, show that they are only marginally better.

While the goal of this paper is not to find the posterior means and triple-goal estimates under the best possible working model, a good working model is expected to improve on both the standard Bayesian and triple-goal methods. Thus model selection will be a problem of great interest before implementing the triple-goal method for producing a new unemployment rate series for multi-purpose uses. In the future, we plan to develop a benchmarked triple-goal estimation system using the multi-level time series model used by BLS for its production of official statistics for the states. Neither of the two methods of evaluation considered in the paper should be considered an ideal method, which does not seem to exist in small area estimation evaluation. But nonetheless our evaluation study should shed some light on the merit of triple-goal for multi-purpose inferences and should encourage researchers to think of new ideas for evaluating small area methods.

Acknowledgements

Authors are listed in alphabetical order. The research of the first and last authors has been supported by the U.S. Census Bureau Prime Contract No: YA1323-09-CQ-0054 (Subcontract No: 41-1016588). The research of the third author has been partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Authors would like to thank Professor Thomas Louis for some useful discussions on the triple-goal method. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the U.S. Census Bureau.

REFERENCES

CONLON, E. M. and LOUIS, T. A. (1999). Addressing Multiple Goals in Evaluating Region-specific Risk Using Bayesian Methods. In *Disease Mapping and Risk Assessment for Public Health*, Chichester: Wiley, 31–47.

DEVINE, O. J. and LOUIS, T. A. (1994). A Constrained Empirical Bayes Estimator for Incidence Rates in Areas with Small Populations. *Statistics in Medicine*, 13, 1119–1133.

EFRON, B. and MORRIS, C. (1975). Data Analysis Using Stein's Estimator and Its Generalizations. *Journal of the American Statistical Association*, 70, 311–319.

GELMAN, A. and PRICE, P. N. (1999). All Maps of Parameter Estimates are Misleading. *Statistics in Medicine*, 18, 3221–3234.

GHOSH, M. (1992). Constrained Bayes Estimation with Applications. *Journal of the American Statistical Association*, 87, 533–540.

GOLDSTEIN, H. and SPIEGELHALTER, D. J. (1996). League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of Royal Statistical Society, Series A.*, 159, 385–409.

GURNEY, M. and DALY, J. F. (1965). A Multivariate Approach to the Estimation in Periodic Sample Surveys. In *Proceedings of the Social Statistics Section, ASA.*, 242–257.

HA, N. S., LAHIRI, P., and PARSONS, P. (2014). Methods and Results for S-mall Area Estimation Using Smoking Data from The 2008 National Health Interview Survey. *Statistics in Medicine*, 33, 3932–3945.

JIANG, J. and LAHIRI, P. (2006). Mixed Model Prediction and Small Area Estimation. *Test*, 15, 111–999.

LAHIRI, P. (1990). Adjusted Bayes and Empirical Bayes Estimation in Population Sampling. *Sankhya*, 52, 50–66.

LAIRD, N. M. and LOUIS, T. A. (1989). Empirical Bayes Ranking Methods. *Journal of Educational Statistics*, 14, 29–46.

LANDRUM, M. B., BRONSKILL, S. E., and NORMAND, S. (2000). Analytic Methods for Constructing Cross-Sectional Profiles of Health Care Providers. *Health Services Outcomes Research Methodology*, 1, 23–47.

LENT, J., MILLER, S., CANTWELL, P., and DUFF, M. (1999). Effects of composite weights Current Population Survey. *Journal of Official Statistics*, 15, 431–448.

LIU, B., LAHIRI, P., and KALTON, G. (2014). Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions. *Survey Methodology*, 40, 1–13.

LOUIS, T. (1984). Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods. *Journal of the American Statistical Association*, 79, 393–398.

PFEFFERMANN, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, 28, 1, 40–68

PFEFFERMANN, D. and TILLER, R. B. (2006). Small Area Estimation With State-Space Models Subject to Benchmark Constraints. *Journal of the American Statistical Association*, 101, 1387–1397.

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

SHEN, W. and LOUIS, T. (1998). Triple-Goal Estimates in Two-Stage Hierarchical Models. *Journal of Royal Statistical Society, Series B.*, 60, 455–471.

STATISTICS IN TRANSITION new series and SURVEY METHODOLOGY
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 523–540

COVARIATE SELECTION FOR SMALL AREA ESTIMATION IN REPEATED SAMPLE SURVEYS

Jan A. van den Brakel[1], Bart Buelens[2]

ABSTRACT

If the implementation of small area estimation methods to multiple editions of a repeated sample survey is considered, then the question arises which covariates to use in the models. Applying standard model selection procedures independently to the different editions of the survey may identify different sets of covariates for each edition. If the small area predictions are sensitive to the different models, this is undesirable in official statistics since monitoring change over time of statistical quantities is of utmost importance. Therefore, potential confounding of true change and methodological alterations should be avoided. An approach to model selection is proposed resulting in a single set of covariates for multiple survey editions. This is achieved through conducting covariate selection simultaneously for all editions, minimizing the average of the edition-specific conditional Akaike Information Criteria. Consecutive editions of the Dutch crime victimization survey are used as a case study. Municipal estimates of three survey variables are obtained using area level models. The proposed averaging strategy is compared to the standard method of considering each edition separately, and to an elementary approach using covariates selected in the first edition. Resulting models, point estimates and MSE estimates are analyzed, indicating no substantial adverse effects of the conceptually attractive averaging strategy.

Key words: area level models, cAIC, Hierachical Bayesian predictors.

1. Introduction

At national statistical institutes, estimation procedures for surveys based on probability samples are traditionally based on design-based or model-assisted inference procedures. Well-known examples are the π-estimator (Narain, 1951; Horvitz and Thompson, 1952) and the generalized regression estimator (Särndal, Swensson and Wretman, 1992). These approaches are particularly appropriate in the case of large sample sizes. In the case of small sample sizes, however, design-based and model-assisted estimators have unacceptably large variances. This occurs when estimates

[1] Statistics Netherlands, Department of Statistical Methods and Maastricht University, Department of Quantitative Economics. E-mail: jbrl@cbs.nl

[2] Statistics Netherlands, Department of Statistical Methods. E-mail: bbus@cbs.nl

are required for detailed breakdowns of the population in subpopulations or domains according to various socio-demographic or geographic classification variables. In such cases, model-based estimation procedures are required to increase the effective sample size of the separate domains with sample information observed in other domains or preceding periods. This class of estimation procedures is known in the literature as small area estimation (SAE) (Rao, 2003; Pfeffermann, 2013) and offers promising opportunities for official statistics (Boonstra et al., 2008).

A common approach to introducing SAE in an existing survey is to apply SAE methods to historic editions of the survey, producing small area estimates for multiple past editions at the same time. This article focuses on the selection of covariates to be used in the SAE models in this setting. In the literature, model selection procedures mostly focus on the selection of optimal models for one particular survey data set (Claeskens and Hjort, 2008). If in each edition of a repeated survey a separate and different model is selected, the question arises to what extent the small area predictions are comparable over time. In official statistics potential confounding of estimates of change over time of some statistic with variations in the inference procedures must be avoided. This article contributes to the existing literature by addressing the question how to select a single optimal model for the production of SAE predictions for independent, repeated editions of a sample survey.

An approach is proposed in which the model selection criterion is averaged over all available editions, leading to a single set of covariates to be used in each edition. This novel approach is compared to the standard approach of selecting a set of covariates for each edition independently using four past editions of the Dutch crime victimization survey. In addition, a simple scenario is included whereby covariates are selected using only the first of a series of survey editions. In this paper models are considered that only use cross-sectional correlation. Alternative approaches that combine cross-sectional and temporal data are proposed by Rao and Yu (1994), Datta et al. (1999) and Pfeffermann and Tiller (2006). These approaches might also be considered to select one single optimal model for subsequent survey editions. These approaches are not considered for implementation in the Dutch crime victimization survey since they are considerably more complex and computationally intensive.

The article continues in Section 2 with a presentation of the SAE methods used and details covariate selection procedures. Section 3 introduces the crime victimization survey and potential covariates. Results are presented and discussed in Section 4. Conclusions are drawn in Section 5.

2. Methods

2.1. Small Area Estimation

In small area estimation multilevel models are used to improve the estimation of small domain parameters. These models use relevant auxiliary information as covariates. In this article the area level model is used (Fay and Herriot, 1979), where the input data for the model are the direct estimates for the domains. Approaches to covariate selection discussed below can be applied to unit level models (Battese, Harter and Fuller, 1988) as well. The area level model is considered, since it takes the complexity of the sample design into account as the dependent variables of the model are the design-based estimates derived from the probability sample and available auxiliary information used in the weighting model of the generalized regression (GREG) estimator. Let $\hat{\theta}_i$ denote the GREG estimates of the target variables θ_i for the domains $i = 1, \ldots, m$. In the area level model, the direct domain estimates are modeled with a measurement error model, i.e. $\hat{\theta}_i = \theta_i + e_i$, where e_i denotes the sampling error with design variance ψ_i. The unknown domain parameter is modeled with available covariates for the $i-$th domain, i.e. $\theta_i = z_i^{'}\beta + v_i$, with z_i a K-vector with the covariates $z_{i,k}$ for domain i, β the corresponding K-vector with fixed effects and v_i the random area effects with variance σ_v^2. For each variable a separate univariate model is assumed. Combining both components gives rise to the basic area level model, originally proposed by Fay and Herriot (1979):

$$\hat{\theta}_i = z_i^{'}\beta + v_i + e_i, \tag{1}$$

with model assumptions

$$v_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2) \text{ and } e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \psi_i). \tag{2}$$

It is assumed that v_i and e_i are independent and that ψ_i is known.

Model(1) is a linear mixed model and estimation often proceeds using Empirical Best Linear Unbiased Prediction (EBLUP), where the between domain variance σ_v^2 is estimated with the Fay-Herriot moment estimator, maximum likelihood or restricted maximum likelihood, see Rao (2003), ch. 6 for details. A weakness of these methods is that in some situations the estimated model variance tends to zero, see e.g. Bell (1999) and Rao (2003). To avoid these problems, the Hierarchical Bayesian (HB) approach is followed in this article, Rao (2003), section 10.3. Therefore, the basic area level model is expressed as an HB model by (1) and (2) and a flat prior on β and σ_v^2. The HB estimates for θ_i and its MSE are obtained as the posterior mean and variance of θ_i. To account for the uncertainty in the between domain variance, integration over the posterior density for σ_v^2 is conducted.

Estimates for the design variances ψ_i are available from the GREG estimator but are used as if the true design variances are known, which is a standard assumption in small area estimation. Therefore, it is important to provide reliable estimates for ψ_i. The stability of the estimates for ψ_i is improved using the following ANOVA-type pooled variance estimator

$$\begin{aligned} \psi_i &= \frac{1-f_i}{n_i} S_p^2, \\ S_p^2 &= \frac{1}{n-m} \sum_{i=1}^{m} (n_i - 1) S_{i;GREG}^2, \end{aligned}$$

with f_i the sample fraction in domain i, n_i the sample size in domain i, $n = \sum_{i=1}^{m} n_i$ and $S_{i;GREG}^2$ the estimated population variance of the GREG residuals.

2.2. Conditional AIC

The model selection procedures discussed here are optimization routines, minimizing the conditional Akaike Information Criterion (cAIC) proposed by Vaida and Blanchard (2005).

The cAIC is applicable to mixed models where the focus is on prediction at the level of clusters or areas (Vaida and Blanchard, 2005). It is defined as cAIC $= -2\mathscr{L} + 2p$, where $\mathscr{L}$ is the conditional log-likelihood and p a penalty based on a measure for the model complexity. In the case of a fixed effects model, p is the number of model parameters. The random part of a mixed model also contributes to the number of model degrees of freedom p with a value between 0 in the case of no domain effects (i.e. $\hat{\sigma}_v^2 = 0$) and the total number of domains m in the case of fixed domain effects (i.e. $\hat{\sigma}_v^2 \to \infty$). In the expression of the cAIC, p is the effective degree of freedom of the mixed model and is defined as the trace of the hat matrix H, which maps the observed data to the fitted values, i.e. $\hat{y} = Hy$, see Hodges and Sargent (2001).

When comparing models, the one with the lowest cAIC value is preferred.

2.3. Covariate selection procedures

Covariate selection procedures are aimed at establishing a set of covariates – in the present setting the fixed effects – to use in models specified by equation (1). This boils down to finding an optimal subset from a larger set of available candidates. All three methods detailed below proceed along the same lines: they follow a step-forward covariate selection strategy which starts from an intercept-only model adding covariates one-by-one until there is no improvement in terms of the selection criterion. This may result in sub-optimal models as the procedure converges to a local minimum of the selection criterion but not necessarily to the global minimum (Claeskens and Hjort, 2008). The focus here, however, is on establishing a single set of covariates for use in repeated survey editions. Alternative search routines converging to the global minimum of the selection criterion can be applied analogously to the step-forward routine used here.

Some general notation is introduced. When C candidate covariates are available for inclusion as a fixed effect in a model specified by equation (1), the set of selected covariates is denoted by s and the set of remaining covariates by r. For ease of use the candidate covariates are assumed to be ordered in a fixed but arbitrary order, so that they can be referred to by their index. For example, a model containing the jth covariate – with $1 \leq j \leq C$ – as a fixed effect, can be identified by $s = \{j\}$. Consequently, in such case $r = \{i\}_{i \neq j}$. Evidently, the equality $s \cup r = \{1, \ldots, C\}$ always holds. Sets of selected and remaining covariates that are specific to a survey edition t are denoted by s_t and r_t respectively.

2.3.1 Selecting an optimal set for each edition separately

For a series of independent cross sectional surveys repeated at times $t = 1, \ldots, T$, a standard covariate selection routine consists in selecting covariates for each edition indepentently.

Covariate selection procedure '*stnd*'. Repeat for all editions $t \in \{1, \ldots, T\}$:

```
Initialization  Set $r_t = \{1, \ldots, C\}$ and $s_t = \{\}$, obtain the corresponding cAIC
    value and call this $cAIC_0$. Set $i = 0$.
Repeat  Attempt extending the model with one covariate:
    a / Set $i = i + 1$.
    b / Calculate cAIC for all models $s_t \cup \{j\}, \forall j \in r_t$, and call these $cAIC_j$.
    c / If $min(cAIC_j) < cAIC_{i-1}$ then set $cAIC_i = min(cAIC_j)$, extend $s_t$ to
        include the corresponding covariate $j$, remove that covariate from
        $r_t$.
Until  The model is not extended or all candidate covariates are included in
    the model.
```

The result are sets s_t of selected covariates for each edition t. In general, s_t and $s_{t'}$ can be different for $t \neq t'$.

The generic model specification given by equation (1) is adapted to reflect the repeated nature of the survey.

$$\hat{\theta}_{i,t} = z_{i,t}^{[stnd]'} \beta_t + v_{i,t} + e_{i,t}, \tag{3}$$

for $i = 1, \ldots, m$ and $t = 1, \ldots, T$, with model assumptions

$$v_{i,t} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{v,t}^2) \text{ and } e_{i,t} \overset{\text{ind}}{\sim} \mathcal{N}(0, \psi_{i,t}). \tag{4}$$

The vectors $z_{i,t}^{[stnd]}$ consist of covariates contained in s_t at the level of the domains i, with s_t established through the *stnd* covariate selection procedure.

2.3.2 Selecting one optimal set for all editions simultaneously

Since the standard method may result in different sets of covariates for different survey editions, an alternative is proposed here, resulting in a single set of covariates for all editions. Formally, the following procedure enforces that $s_t = s_{t'}$ for all $t, t' \in \{1, \ldots, T\}$.

Covariate selection procedure '*avrg*'. Consider all survey editions $t = 1 \ldots T$ simultaneously.

```
Initialization Let $r = \{1, \ldots, C\}$ and $s = \{\}$. Use $r$ and $s$ for all $t$, obtain
    the corresponding cAIC values, and call these $cAIC_{0,t}$. Define $cAIC_0 = \frac{1}{T}\sum_t cAIC_{0,t}$. Set $i = 0$.
Repeat Attempt extending the model with one covariate:
    a / Set $i = i + 1$.
    b / For all editions $t \in \{1, \ldots, T\}$, calculate cAIC for all models $s \cup \{j\}$,
        $\forall j \in r$, and call these $cAIC_{j,t}$.
    c / Define $cAIC_j = \frac{1}{T}\sum_t cAIC_{j,t}$.
    d / If $min(cAIC_j) < cAIC_{i-1}$ then set $cAIC_i = min(cAIC_j)$, extend $s$ to
        include the corresponding covariate $j$, remove that covariate from
        $r$.
Until The model is not extended or all candidate covariates are included in
    the model.
```

This strategy is based on averaging the model selection criterion cAIC and results in a single set s of covariates to be used in all editions t. The corresponding model specification, with a fixed set of covariates for repeated surveys, is written as (3) and (4) where the vectors $z_{i,t}^{[stnd]}$ are replaced by vectors, say $z_{i,t}^{[avrg]}$, that consist of covariates contained in s at the level of the domains i at the time periods t, with s established through the *avrg* covariate selection procedure.

2.3.3 Selecting an optimal set based on the first edition only

An elementary approach also resulting in a single set of covariates is to use the first edition of a series of repeated surveys to establish the set of covariates and to use these in all subsequent editions.

Covariate selection procedure '*frst*'.

Apply procedure *stnd* for $t = 1$ to obtain s_1.

The set of covariates s_1 obtained based on the first edition is used at all times. The model takes the form of (3) and (4) where the vectors $z_{i,t}^{[stnd]}$ are replaced by vectors, say $z_{i,t}^{[frst]}$, that consist of covariates contained in s_1 at the level of the domains i at the time periods t. This strategy is included to assess and illustrate its performance. In other settings than the one discussed in the present article, statisticians may be in a situation where a survey is foreseen to be repeated in the future, but SAE estimates are required at the time of the first edition. The only option then is to use that edition for covariate selection.

3. Data

3.1. Crime victimization survey

The Dutch crime victimization survey underwent several redesigns in the past, including in 2008 and 2012. In the period from 2008 through 2011 the survey is known as the Integrated Safety Monitor (ISM). These four editions of the ISM are used as a case study in the present article. The purpose of the ISM is to publish information on crime victimization, public safety and satisfaction with police performance, among others. Each annual ISM sample is obtained independently through stratified simple random sampling of persons aged 15 years or older residing in the Netherlands. The population register serves as the sampling frame. The country is divided into 25 police districts, which are used as the stratification variable in the sample design. The yearly sample size of about 19,000 respondents is divided equally over the strata. In addition to this national sample, local authorities such as municipalities and police districts can draw supplementary samples in their own regions on a voluntary basis, with the purpose to obtain precise local estimates.

These supplementary samples are also based on stratified simple random sampling, but now with a more detailed geographical stratification variable, usually neighborhood. Table 1 gives an overview of the oversampling and the number of respondents for the years 2008 through 2011. Participation in the oversampling scheme by local authorities was encouraged in the years 2009 and 2011 resulting in much larger samples in these editions.

Table 1: Overview of response and oversampling in ISM surveys 2008 - 2011.

	2008	2009	2010	2011
Number of oversampled municipalities	77	239	21	225
Size response national sample	16,964	19,202	19,238	20,325
Size response supplemental sample	45,839	182,012	19,982	203,621
Percentage of population in oversampled areas	29%	65%	16%	66%

Data collection is based on a sequential mixed mode design using internet (WI), paper (PAPI), telephone interviewing (CATI) or face-to-face interviewing (CAPI). For the data collection of the additional regional samples the WI, PAPI and CATI modes are mandatory. The use of the CAPI mode is recommended but not mandatory since this mode is very costly. Statistical inference for official publication purposes is based on the GREG estimator. The inclusion probabilities in the ISM are determined by the sampling design, accounting for stratification and oversampling at regional levels. The GREG estimator uses a complex weighting scheme that is based on the auxiliary variables age, gender, ethnicity, urbanization, household size, police district, and the strata used in the regional oversampling scheme. In addition, the weighting scheme contains a component that calibrates the response to a fixed distribution over the data collection modes with the purpose to stabilize the measurement error between the subsequent editions of the ISM, (Buelens and Van den Brakel, 2015). Variance estimates are obtained with the standard Taylor series approximation of the GREG estimator, see Särndal, Swensson and Wretman (1992), ch. 6.

The GREG estimator can be used to produce reliable official statistics for regions with relatively large sample sizes. With the aforementioned sample design this implies that the GREG estimator can be used to produce official statistics at the level of police districts and in the regions where additional samples are drawn also at the level of municipalities. For regions where no additional samples are drawn, sample sizes are too small to produce reliable estimates at the level of municipalities with the GREG estimator. Since there is a growing demand for such figures, SAE procedures are developed to produce reliable official statistics on crime victimization at the municipal level. Three important ISM variables under study are listed in

Table 2.

Table 2: Overview of key ISM variables and their associated statistics.

Variable	Description of statistic
victim	Percentage of people who indicated that they were a victim of crime in the last 12 months
degen	Degeneration of the neighborhood (on a scale 1-5)
contpol	Percentage of people who had contact with the Police in the last 12 months

3.2. Candidate covariates

The success of increasing the precision of the domain estimates using SAE methods critically depends on the availability of correlated auxiliary information. An overview of 21 potential covariates used for model building is given in Appendix A. These are obtained from the Police Register of Reported Offences (PRRO) and from the population register.

The auxiliary variables `mode2` and `oversampled` require some explanation. In areas where local authorities draw supplemental samples, the fraction of responses obtained through non-interviewer administered modes is larger compared to areas without such oversampling. This is caused by the fact that CAPI is not conducted for the supplemental samples. There are clear indications that there are systematic differences in measurement error between responses obtained through interviewer and non-interviewer administered modes (Buelens and Van den Brakel, 2015; Schouten et al., 2013). As mentioned in section 3.1, the GREG estimator calibrates the response to fixed mode distributions to level out large fluctuations in measurement error due to large fluctuations in the distribution of the response over the different modes (Buelens and Van den Brakel, 2015). Since the calibration occurs at the police district level and not at the municipal level, it can be expected that the fraction of non-interviewer administered modes or a dummy indicator to differentiate between municipalities where oversampling took place or not, has predictive power for at least some of the target variables, due to potential correlation between these covariates and mode-dependent measurement error present within the municipal estimates.

4. Results

The different covariate selection strategies are applied to the four ISM editions for selecting covariates for SAE models for the three study variables. The sections below discuss the sets of selected covariates and compare performance of the resulting models. The HB estimates are computed using the statistical software environment R (R Development Core Team, 2009) and package `hbsae` (Boonstra, 2012).

4.1. Covariate selection results

The covariate selection results are given in Tables 3 and 4. When using the *stnd* approach, different sets of covariates are selected in different survey editions. Not only do the covariates differ, also their number can vary between years. The variables selected through the *avrg* strategy often appear in at least one of the *stnd* models. The *frst* approach is not listed in Table 3 as it uses the set of covariates selected through the *stnd* approach in 2008.

Naturally, the *stnd* models result in lower cAIC values than the other strategies, see Table 4. By definition, the *avrg* and *frst* procedures result in the same sets of covariates to be used for all editions. For 2008, the *frst* and *stnd* approaches are identical and can therefore be expected to perform better than the *avrg* approach in that edition. For the subsequent years, 2009-2011, the cAIC values associated with the *avrg* approach are mostly smaller than or equal to the cAIC values obtained with the *frst* approach. In some cases the covariates selected for 2008 perform well in other years too, this is the case for example with `victim` in 2010.

4.2. Small area estimates

The purpose of applying SAE techniques in official statistics is to increase precision of area estimates. When considering the use of the *avrg* or *frst* approaches it is of interest to compare the reductions in variance achieved with these strategies compared to the *stnd* approach. An appropriate quantity to study in this context is the mean reduction in the coefficient of variation (MRCV),

$$MRCV = \frac{1}{m}\sum_{i=1}^{m} \frac{CV(\hat{\theta}_{i,t}) - CV(\tilde{\theta}_{i,t})}{CV(\tilde{\theta}_{i,t})}, \tag{5}$$

with $\hat{\theta}_{i,t}$ the GREG estimator and $\tilde{\theta}_{i,t}$ the HB prediction for domain i, and $CV(x)$ the coefficient of variation of estimator x (the estimated standard error divided by the point estimate). Note that MRCV would not be a suitable model selection criterion as it is susceptible to over fitting.

Table 3: Covariates selected by the different methods. Strategy *frst* uses the covariates selected for *stnd* in 2008.

Variable	Method	Covariates (listed in order in which they were selected)
victim	stnd 2008	logdens, propcrimedef2, oversampled, nonwestimmi, old, westimmi, highincome, lowincome, density, carsphh
	stnd 2009	sqrtdens, propcrimedef2, carsphh, mode2, old, westimmi
	stnd 2010	sqrtdens, propcrimedef1, young
	stnd 2011	propcrimedef1, sqrtdens, old, totcrime, mode2, rent, nonwestimmi, meanvalue, lowincome, oversampled
	avrg	sqrtdens, propcrimedef1, young, oversampled, totcrime, westimmi
degen	stnd 2008	rent, totcrime, prov, old, meanvalue, unemployed
	stnd 2009	rent, prov, totcrime, meanvalue, mode2, old, young
	stnd 2010	rent, prov, meanvalue, violcrime, oversampled, mode2, totcrime, biketheft, old, density, logdensity, lowincome
	stnd 2011	rent, totcrime, biketheft, mode2, old, meanvalue, logdens, violcrime, lowincome, carsphh
	avrg	rent, prov, violcrime, meanvalue, totcrime, mode2, biketheft, oversampled, old, propcrimedef2
contpol	stnd 2008	logdens
	stnd 2009	sqrtdens, violcrime, young, mode2
	stnd 2010	logdens, westimmi
	stnd 2011	logdens, violcrime, biketheft, westimmi, prov, highincome
	avrg	logdens, violcrime, westimmi

The MRCV values obtained in this study are listed in Table 5. While the largest reductions are naturally achieved with the *stnd* models, the suboptimality of the *avrg* and *frst* models is mild. Overall, the reductions achieved with the latter methods are only a few percentage points smaller than those achieved with the optimal models. Comparing the *avrg* and *frst* approaches, the former mostly result in greater reductions, although not always.

Table 4: Covariate selection results.

		number of covariates				cAIC			
Variable	Method	2008	2009	2010	2011	2008	2009	2010	2011
victim	stnd	10	6	3	10	-949	-1308	-914	-1395
	avrg	6	6	6	6	-934	-1308	-902	-1381
	frst	10	10	10	10	-949	-1304	-905	-1393
degen	stnd	6	7	12	10	724	353	761	273
	avrg	10	10	10	10	722	357	766	276
	frst	6	6	6	6	724	361	784	294
contpol	stnd	1	4	2	6	-161	-661	-698	-811
	avrg	3	3	3	3	-157	-657	-696	-805
	frst	1	1	1	1	-161	-657	-693	-798

Table 5: Mean reduction in coefficient of variation (in %).

Variable	Method	2008	2009	2010	2011
victim	stnd	-76	-57	-88	-63
	avrg	-72	-56	-86	-60
	frst	-76	-55	-83	-60
degen	stnd	-47	-31	-55	-32
	avrg	-46	-31	-53	-32
	frst	-47	-31	-50	-30
contpol	stnd	-92	-86	-83	-82
	avrg	-90	-86	-83	-87
	frst	-92	-87	-82	-88

Comparing Tables 4 and 5, it is observed that cAIC and MRCV values do not always exhibit the same pattern. For example the cAIC values for the variable `contpol` in 2011 indicate that the *stnd* approach is best, followed by the *avrg* and *frst* approaches. The corresponding MRCV values on the other hand reverse this pattern, with the *frst* approach resulting in greatest reduction and the *stnd* in lowest.

The values in Table 5 indicate that the SAE method in this case is most beneficial for the variable `contpol` with reductions in the coefficient of variation of up to almost 90%. The gains in precision for `victim` are smaller and for `degen` the smallest at around 30% in 2009 and 2011.

In line with the observation in section 3.1 that the oversampling in the ISM was much more intense in 2009 and 2011, it is seen in Table 4 that the cAIC values for these years are smaller than for 2008 and 2010 for each variable and method, indicating better model fits in editions with larger samples. The gains to be had from SAE, however, are larger in the editions with smaller sample sizes, see Table 5.

Of practical relevance is the effect of the covariate selection strategy on the HB point estimates. SAE estimates obtained through the *stnd*, *avrg* and *frst* approaches are compared to GREG estimates in Figure 1. Four municipalities with varying sample sizes are chosen as an example. The number at the top of each panel in the plot refers to the rank of the municipality when ordered according to sample size (0001 being the smallest, and 0418 the largest). The four types of estimates are compared. The differences between the three types of SAE estimates are much smaller than the difference between the SAE and GREG estimates, apart from the larger municipalities where all estimates almost coincide, such as in Amsterdam. In the smaller municipalities, where the sample sizes are generally smaller, the differences are larger and the advantage of using SAE methods becomes apparent. While the *avrg* and *frst* approaches lead to point estimates close to those obtained through the *stnd* approach, sometimes there are differences, in particular in the smaller municipalities. An example is the variable `degen` in municipality '0008' (top left in middle panel of Fig. 1). There, the *avrg* estimates are closer to the *stnd* estimates than the *frst* estimates for the years 2009-2011. This is an indication that situations can arise where the covariates selected in 2008 are suboptimal in later editions, while those selected through the averaging strategy perform better overall.

More detailed results are available online in a Statistics Netherlands research report (Buelens and Van den Brakel, 2014). In this document the point estimates and variance estimates under the different model selection procedures are compared. Additional information on model evaluation is also included in this paper.

5. Conclusion

The issue considered in this article is the choice of model covariates when applying small area estimation repeatedly in consecutive, independent editions of a survey. The model under consideration is the area level model known as the Fay-Herriot model in combination with an Hierarchical Bayesian prediction approach. Model selection in this setting boils down to selecting an optimal set of covariates from a set of possible candidates.

While selecting an optimal set of covariates for each edition separately may be preferable from a modeling perspective, in official statistics it is important to avoid all potentially confounding elements in estimation of temporal change of published statistical results. Using the same set of covariates in SAE models every year is deemed essential. A strategy is proposed in which all editions of a survey are considered simultaneously, and a single set of covariates is selected. This approach uses the cAIC criterion and operates by minimizing the cAIC averaged over all survey editions. A simple additional approach is included in the analyses, consisting of selecting covariates based on the first edition of a survey and using this set in all subsequent editions.

In the four editions of the crime victimization survey, it is shown that the models obtained through the averaging approach are only mildly suboptimal. The resulting coefficients of variation are marginally larger than those obtained for estimates based on specific optimal models for each edition. Models based on the first edition only are somewhat worse than the models obtained through averaging, but not substantially. In this application, point estimates are found to be very similar under all three SAE approaches, with the estimates obtained through the averaging models closer to the optimal models than the estimates obtained by using only the first edition. The models obtained through the averaging approach are used to produce official statistics about crime victimization and public safety at the municipal level, for twelve ISM survey variables in addition to the three discussed in this article.

The fact that using the first edition of a repeated survey to establish models once and that using them unaltered thereafter provides reasonable results not dramatically different from using optimal models in each edition, is an empirical finding for this application. This is the approach that would ordinarily be taken when a new survey is introduced with the plan to repeat it at future points in time. When SAE statistics are required in the first edition there is no other option than to base model selection on that edition alone. When multiple editions are available, however, it is recommendable to conduct model selection on these editions simultaneously using the proposed averaging strategy. Even if a number of past editions are available, it remains necessary to evaluate the selected models if the data under new editions of the survey become available. Changing the model might require a revision strategy for figures already published in the past.

While the averaging method is developed for area level models in the present study, it is in principle applicable to situations with other models as well including the unit level model (Battese, Harter and Fuller, 1988) and models with spatial effects (You and Zhou, 2011). Similarly, other model selection criteria than cAIC could be used if desired. Buelens and Van den Brakel (2014) considered leave-one-out cross validation and found it to result in less parsimonious models than cAIC. Recently, Lahiri and Suntornchost (2015) proposed a new variable selection criterion specifically for Fay-Herriot models. Each of these alternatives can immediately be plugged into the selection strategies presented in this article.

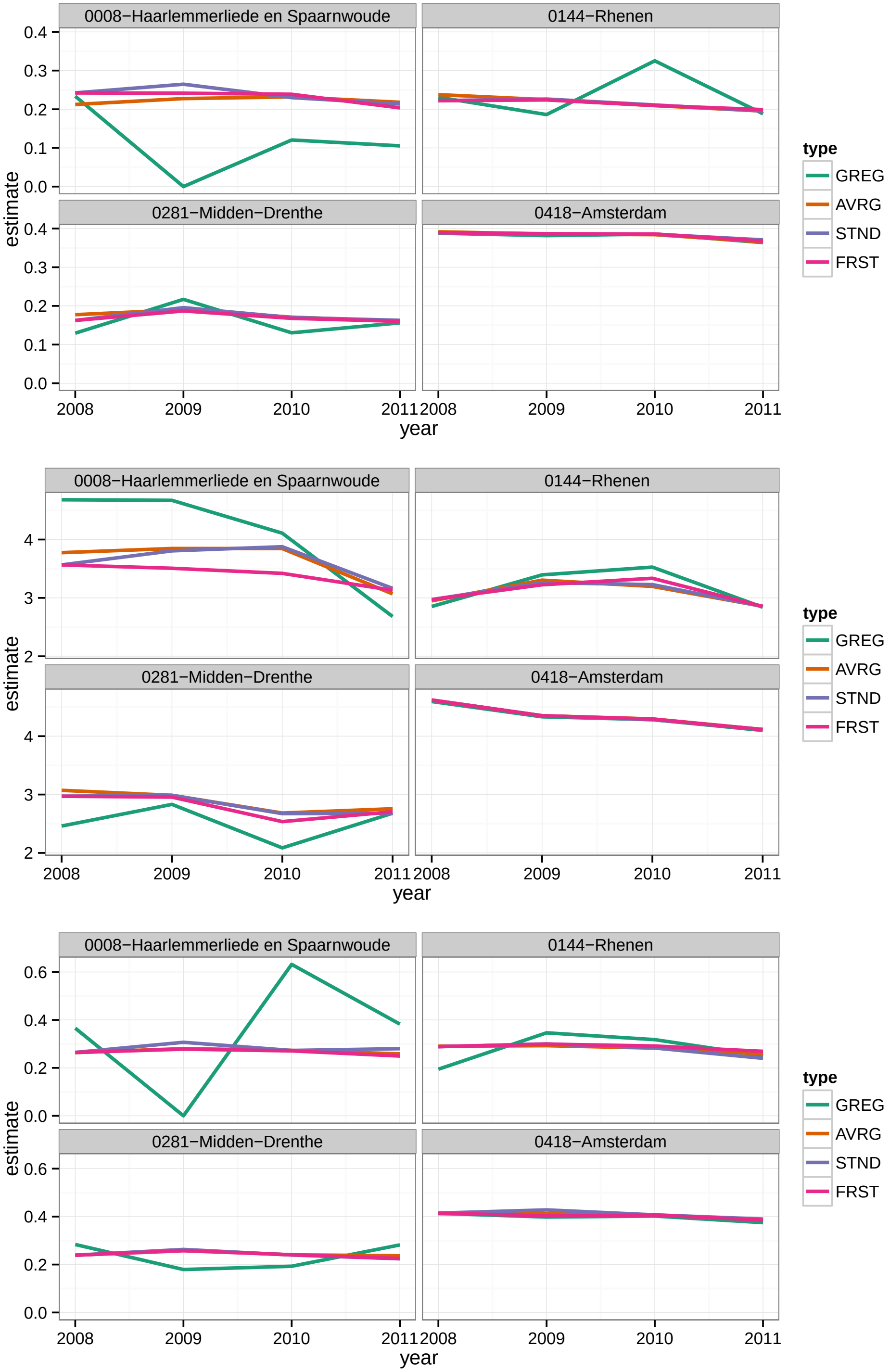

Figure 1: Time series of GREG and SAE estimates obtained through the *stnd*, *avrg* and *frst* approaches for four municipalities for `victim` (top), `degen` (middle) and `contpol` (bottom).

Appendix A: Auxiliary variables defined for municipalities

westimmi:	share of western immigrants in the population
nonwestimmi:	share of non-western immigrants in the population
prov:	province
density:	housing density (number of dwellings per square kilometer)
logdens:	natural logarithm of density
sqrtdens:	square root of density
meanvalue:	mean house value (available from housing register)
carsphh:	average number of cars owned by households
young:	share of population aged 15-30
old:	share of population aged 65+
rent:	share of houses that are rented (as opposed to owned)
lowincome:	share of households with a low income (nationwide in lowest quintile)
highincome:	share of households with a high income (nationwide in highest quintile)
unemployed:	share of population registered at the employment agency as looking for work
totcrime:	number of crimes registered by the Police per 1.000 inhabitants
propcrimedef1:	number of property crimes registered by the Police per 1.000 inhabitants (definition CBS)
propcrimedef2:	number of property crimes registered by the Police per 1.000 inhabitants (definition Bureau Veiligheid)
biketheft:	number of bicycle thefts registered by the Police per 1.000 inhabitants
violcrime:	number of violent crimes registered by the Police per 1.000 inhabitants
mode2:	share of non-interviewer administered modes (paper and web) in the ISM survey
oversampled:	binary variable indicating whether the municipality took part in the ISM oversampling scheme

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, 28–36.

BELL, W.R., (1999). Accounting for uncertainty about variances in small area estimation. *Technical report*. Bulletin of the International Statistical Institute.

BOONSTRA, H.J., (2012). hbsae: Hierarchical Bayesian Small Area Estimation, *Manual R package version 1.0.*. Statistics Netherlands, Heerlen.

BOONSTRA, H.J., VAN DEN BRAKEL, J.A., BUELENS, B., KRIEG, S., SMEETS, M., (2008). Towards small area estimation at Statistics Netherlands. *METRON International Journal of Statistics*, LXVI, 21–49.

BUELENS, B., VAN DEN BRAKEL, J.A., (2014). Model selection for small area estimation in repeated surveys. *Discussion paper 201423*, Statistics Netherlands, Heerlen. http://www.cbs.nl/NR/rdonlyres/308ED398-714A-41A4-A57 C-9DCCC3F30D35/0/201423x10pub.pdf

BUELENS, B., VAN DEN BRAKEL, J.A., (2015). Measurement error calibration in mixed mode surveys. *Sociological Methods and Research*, 44, 391–426.

CLAESKENS, G., HJORT, N.L., (2008). *Model selection and model averaging*, Cambridge series on statistical and probabilistic mathematics, Cambridge University Press.

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

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, 268–277.

HODGES, J.S., SARGENT, D.J., (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika*, 88, 367–379.

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.

LAHIRI, P., SUNTORNCHOST, J., (2015). Variable selection for linear mixed models with applications in small area estimation. *Sankhya B*, 1–9, doi = 10.1007/s13571-015-0096-0.

NARAIN, R., (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 581–613.

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

PFEFFERMANN, D., TILLER, R., (2006). Small Area Estimation with State-Space Models subject to Benchmark Constraints. *Journal of the American Statistical Association*, 101, 1387–1397.

R DEVELOPMENT CORE TEAM, (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

RAO, J.N.K., (2003). *Small Area Estimation*, New York: John Wiley.

RAO, J.N.K., YU, M., (1994). Small-area estimation by combining time-series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.

SÄRNDAL, C-E., SWENSSON, B., WRETMAN, J., (1992). *Model Assisted Survey Sampling*, New York: Springer.

SCHOUTEN, B., VAN DEN BRAKEL, J.A., BUELENS, B., VAN DER LAAN, J., KLAUSCH, T., (2013). Disentangling mode-specific selection bias and measurement bias in social surveys. *Social Science Research*, 42, 1555-1570.

VAIDA, F., BLANCHARD, S., (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, 92, 351–370.

YOU, Y., ZHOU, Q., (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology*, 37, 25–36.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 541–562

SMALL AREA ESTIMATION FOR SKEWED DATA IN THE PRESENCE OF ZEROES

Forough Karlberg[1]

ABSTRACT

Skewed distributions with representative outliers pose a problem in many surveys. Various small area prediction approaches for skewed data based on transformation models have been proposed. However, in certain applications of those predictors, the fact that the survey data also contain a non-negligible number of zero-valued observations is sometimes dealt with rather crudely, for instance by arbitrarily adding a constant to each value (to allow zeroes to be considered as "positive observations, only smaller", instead of acknowledging their qualitatively different nature).

On the other hand, while a lognormal-logistic model has been proposed (to incorporate skewed distributions as well as zeroes), that model does not include any hierarchical aspects, and is therefore not explicitly adapted to small area prediction.

In this paper, we consolidate the two approaches by extending one of the already established log-transformation mixed small area prediction models to incorporate a logistic component. This allows for the simultaneous, systematic treatment of domain effects, outliers and zero-valued observations in a single framework. We benchmark the resulting model-based predictors (against relevant alternatives) in applications to simulated data as well as empirical data from the Australian Agricultural and Grazing Industries Survey.

Key words: small area estimation, representative outliers, zero-valued observations, lognormal-logistic mixture model.

1. Introduction

1.1. Estimation in the presence of skewed data

It is a well-known fact that survey data frequently are skewed (Huber 1981, Fuller 1991, Barnett and Lewis 1994). Examples include the income (Mincer 1970) and wealth (Huggett 1996) of private individuals as well as many of the variables observed in Business surveys (Chambers 1986, Thorburn 1993, Hidiroglou and

[1] Luxembourg Statistical Services. E-mail: Forough.Karlberg@LuxStat.eu.

Smith 2005, Zimmermann and Münnich 2013, Shlomo and Priam 2013). These extreme values are not erroneous; on the contrary, to take but one example, a large enterprise typically constitutes an important part of the local economy of a municipality – and to treat them as anomalies by merely eliminating them when they are encountered would be erroneous. Such extreme values are to be regarded as representative outliers in the terminology of Chambers (1986). Various methods have been developed to treat the issue of estimation in the presence of such outliers, e.g. by adjusting outlyingness, possibly in connection with determining a boundary (threshold) for the outliers (Searls 1966, Kokic 1998, Hubert and Van der Veeken 2007), as well as some methods with downweighting (Hidiroglou and Srinath 1981, Lee 1995, Sinha and Rao 2009). Historically, there are different approaches used for transforming the data (including important outliers) to linearity (Carroll and Ruppert 1988, Chen and Chen 1996, Chandra and Chambers 2011, Berg and Chandra 2012), with some applications concentrating on the finite population distribution of a survey variable (Royall 1982, Jiang and Lahiri 2006, Salvati *et al.*, 2012). Karlberg (2000a) conducts model-based estimation under a lognormal model and extends it to a lognormal-logistic (Karlberg, 2000b). This has the double advantage of moderating the impact of outliers that are in the sample and, in case no outliers are included, to adjust for their (assumed) presence in the population. However, there are also issues with lognormal models. First, the back-transformation introduces bias which must be corrected for; while technically challenging, this is manageable; bias-correction terms are provided by, e.g. Karlberg (2000b). More importantly, as with all model-based estimation, severe bias could result in case the presumed lognormal model does not hold.

By logical extension, small area estimation involving skewed variables is also a challenge, compounded by the fact that the samples for each domain are smaller, leading to an even higher sensitivity to outliers (Lehtonen *et al.*, 2003). Various methods, some of them including log-transformation of the data, have been proposed (Chambers and Dorfman 2003, Slud and Maiti (2006), Chandra and Chambers 2011, Berg and Chandra 2012, Zimmermann and Münnich 2013).

1.2. The added complexity of zero-valued observations

It is not infrequent to encounter skewed variables that, while considerably right-skewed, also contain a sizeable proportion of zero-valued observations (Lamberta 1992, Chen *et al.*, 2003). Obviously, estimation methods based on logarithmic transformation are no longer directly applicable to such variables. Sometimes, this is addressed by merely adding an arbitrary constant κ (κ=1 being common practice) to the variable (see Young and Young 1975), which then again becomes possible to logarithm. However, this manner of treating zero-valued observations is not unproblematic. First, from a technical point of view, it is hard to argue that the resulting logarithmed variable is normally distributed – it would rather be bimodal, with one mode at $\ln(\kappa)$, and definitely not continuous, with a large number of values assuming the exact same value $\ln(\kappa)$. Moreover, the choice of the constant κ is arbitrary, with a different choice rendering different results. Finally, and most importantly, it could be argued that a variable assuming the value 0 is something more than a computational problem or a technical nuisance – sample units with zero-valued observations are in fact often qualitatively different from those with positive values. Taking wages as an example, a person with a wage figure of 0 is typically not "gainfully employed but with a salary of 0", but rather unemployed or otherwise out of the labour market. Similarly, a farm with a crop area of 0 does typically not belong to a crop farmer who just happens to not grow any crops, but rather to a farmer focusing on other activities, such as dairy, forestry or livestock.

1.3. Solutions investigated in this paper

The lognormal-logistic model discussed by Karlberg (2000b) seems to be a more appropriate way to address this issue. The estimator associated with that model first fits a logistic model (to deal with the zero-valued observations), and thereafter fits a lognormal model to the positive observations. However, the model in question is not directly designed to accommodate small area estimation. In this paper, we will therefore devote Section 2 to extending the model of Karlberg to incorporate hierarchical elements (or, put differently, extending the model of Berg and Chandra (2012) to incorporate a logistic element). This is achieved by straightforward, practical combinations of already existing tools (see Pfeffermann, 2013); this paper includes no major theoretical contributions. The empirical properties of the four resulting estimators are then examined in Section 3, for random lognormal-logistic data, as well as for data from the Australian Agricultural and Grazing Industries Survey (AAGIS). The findings are discussed in Section 4, which also brings up possible future lines of study.

2. Methods

2.1. The lognormal-logistic model

Under the lognormal-logistic model studied in this paper, we will, just like Karlberg (2000b), assume that Y_{ij} , the value of unit j for area i for the variable of interest(Y), is the product

$$\mathrm{Y_{ij}} = \widetilde{\mathrm{Y}}_{\mathrm{ij}}\Delta_{\mathrm{ij}}$$

of a "lognormal component" $\widetilde{Y}_{ij}$ and a binary (0 or 1) "logistic component" Δ_{ij} with independence between the two components.

2.1.1. The lognormal component

Letting X_{ij} denote a vector of auxiliary variables for unit j, we assume that

$$\ln(\tilde{Y}_{ij}) = \tilde{Z}_{ij} = \mathbf{BX}_{ij} + u_i + e_{ij}$$

where **B** is an unknown parameter, and, for the area-level effects, we have that they are i.i.d.

$$u_i \sim N(0, \sigma_u)$$

and for the residuals that they are i.i.d.

$$e_{ij} \sim N(0, \sigma_e)$$

with, furthermore, independence between any u_i and any e_{ij}.

2.1.2. The logistic component

Letting Ξ_{ij} denote a vector of auxiliary variables for unit j (possibly identical X_{ij}), we assume that the logistic component values are conditionally independently Bernoulli distributed:

$$\Delta_{ij} \sim \text{Bernoulli}\left(\frac{\exp(\boldsymbol{\beta}\boldsymbol{\Xi}_{ij} + \omega_i)}{1 + \exp(\boldsymbol{\beta}\boldsymbol{\Xi}_{ij} + \omega_i)}\right)$$

where **β** is an unknown parameter and the area-level effects are i.i.d.

$$\omega_i \sim N(0, \sigma_\omega).$$

2.1.3. Relationship with previous models

We see from the first column of Table 1 that estimators for unit-level lognormal models (without a logistic component) have been defined without area effects by Karlberg (2000a) and with area effect by Berg and Chandra (2012). From the two other columns (with stochastic Δ_{ij}), we see, however, that to date, only the simplest case (i.e. with no hierarchical components) has been treated; this corresponds to Karlberg (2000b).

In this paper, we will therefore proceed to investigate lognormal-logistic estimators of small area means corresponding to all four possible cases.

Table 1. Relationship between the model parameters and previously addressed models

	$\Delta_{ij} \equiv 1$ *(i.e. no logistic component)*	Δ_{ij} stochastic	
		$\sigma_\omega = 0$	$\sigma_\omega > 0$
$\sigma_u = 0$	Karlberg (2000a)	Karlberg (2000b)	–
$\sigma_u > 0$	Berg and Chandra (2012)	–	–

2.2. Fitting the model and estimation of small area means

2.2.1. Estimation of the model parameters and fitted area effects

In order to evaluate the various estimators, a simulation study has been conducted. Due to the availability of appropriate SAE packages in R, the study was set up through a couple of R scripts. For all four possible options, the estimation procedure proposed in this paper is as follows:

1. First, the logistic model parameters are estimated. Two cases are possible:

   a. If there is no logistic area effect (i.e. if σ_ω=0), the logistic parameter $\boldsymbol{\beta}$ is estimated by means of logistic regression via the GLM function.

   b. If $\sigma_\omega > 0$, the parameters $\boldsymbol{\beta}$ and σ_ω are estimated (and the ω_i-values are fitted) using hierarchical logistic regression via the HGLM function (Rönnegård *et al.*, 2010).

2. Based on the logistic regression outcome:

   a. Estimated probabilities are computed for each unit as

$$\hat{\mathrm{p}}_{\mathrm{ij}} = \frac{\exp\left(\widehat{\boldsymbol{\beta}}\boldsymbol{\Xi}_{\mathrm{ij}} + \widehat{\omega}_{\mathrm{i}}\right)}{1 + \exp\left(\widehat{\boldsymbol{\beta}}\boldsymbol{\Xi}_{\mathrm{ij}} + \widehat{\omega}_{\mathrm{i}}\right)},$$

   b. area frequencies with positive Y_{ij} values are estimated by

$$\widehat{\mathrm{N}}_{+\mathrm{i}} = \sum\nolimits_{\mathrm{j}\in \mathrm{s_i}} \Delta_{\mathrm{ij}} + \sum\nolimits_{\mathrm{j}\in \mathrm{r_i}} \hat{\mathrm{p}}_{\mathrm{ij}} \text{ , and}$$

   c. area auxiliary variable averages for the observations with positive Y_{ij} values are estimated by

$$\widehat{\bar{\mathbf{X}}}_{+\mathrm{i}} = \left(\sum\nolimits_{\mathrm{j}\in \mathrm{s_i}} \Delta_{\mathrm{ij}}\mathbf{X}_{\mathrm{ij}} + \sum\nolimits_{\mathrm{j}\in \mathrm{r_i}} \hat{\mathrm{p}}_{\mathrm{ij}}\mathbf{X}_{\mathrm{ij}}\right)/\widehat{\mathrm{N}}_{+\mathrm{i}} \,.$$

3. Thereafter, the lognormal model parameters are estimated.

   a. If there is no lognormal area effect (i.e. if σ_u=0), **B** and σ_e are fitted as in Karlberg (2000b).

b. If $\sigma_u > 0$, the parameters **B**, σ_u and σ_e are estimated (and the u_i-values are fitted) as in Battese, Harter and Fuller (1988) using the eblupBHF function (Molina and Marhuenda, 2013), i.e. the empirical best linear unbiased predictor (EBLUP; see Rao 2003, and Wang and Fuller 2003).

2.2.2. Prediction of unobserved values

If there is no lognormal area effect, then the lognormal component of each unobserved value is predicted, as in Karlberg (2000b) by the back-transformed predicted values of Z_{ij} multiplied by a bias correction factor:

$$\hat{\hat{\mathrm{Y}}}_{\mathrm{ij}} = \exp\left(\hat{\hat{\mathrm{Z}}}_{\mathrm{ij}}\right)\exp\left(\frac{\hat{\sigma}_{\mathrm{e}}^{2}}{2}\left(1-\mathrm{a}_{\mathrm{ij}}\right)+\frac{\hat{\sigma}_{\mathrm{e}}^{4}}{4\mathrm{n}_{+}}\right)$$

where n_+ is the number of positive observations in the sample (obtained as the sum of all observed values of Δ_{ij}),

$$\mathrm{a}_{\mathrm{ij}} = \mathbf{X}_{\mathrm{i}}^{\prime}\left(\mathbf{X}^{\prime}\mathbf{X}\right)^{-1}\mathbf{X}_{\mathrm{j}}$$

and

$$\hat{\hat{\mathrm{Z}}}_{\mathrm{ij}} = \hat{\mathbf{B}}\mathbf{X}_{\mathrm{ij}}\ .$$

If the model incorporates lognormal area effects, then the lognormal components are instead predicted, as in Berg and Chandra (2012), by

$$\hat{\hat{\mathrm{Y}}}_{\mathrm{ij}} = \exp\left(\hat{\hat{\mathrm{Z}}}_{\mathrm{ij}}\right)\exp\left(\frac{\hat{\sigma}_{\mathrm{e}}^{2}}{2}\left(\frac{\gamma_{\mathrm{i}}}{\mathrm{n}_{+\mathrm{i}}}+1\right)\right)$$

where the number of positive observations in area i is denoted by

$$\mathrm{n}_{+\mathrm{i}} = \textstyle\sum_{\mathrm{j}\in\mathrm{s}_{\mathrm{i}}}\Delta_{\mathrm{ij}},$$

$$\gamma_{\mathrm{i}} = \hat{\sigma}_{\mathrm{u}}^{2}/(\hat{\sigma}_{\mathrm{u}}^{2}+\hat{\sigma}_{\mathrm{e}}^{2}/\mathrm{n}_{+\mathrm{i}}),$$

and

$$\hat{\hat{\mathrm{Z}}}_{\mathrm{ij}} = \hat{\mathbf{B}}\mathbf{X}_{\mathrm{ij}}+\hat{\mathrm{u}}_{\mathrm{i}}\ .$$

Combining this with the logistic probability estimates, each unobserved value is predicted by

$$\hat{\mathrm{Y}}_{\mathrm{ij}} = \hat{\hat{\mathrm{Y}}}_{\mathrm{ij}}\hat{\mathrm{p}}_{\mathrm{ij}}\ .$$

2.2.3. Estimation of small area means

Finally, based on the sum of the observed and predicted values, the small area means are simply estimated by:

$$\widehat{\overline{\mathrm{Y}}}_{\mathrm{i}} = \frac{1}{\mathrm{n_i}}\left(\sum_{\mathrm{j}\in \mathrm{s_i}} \mathrm{Y_{ij}} + \sum_{\mathrm{j}\in \mathrm{r_i}} \widehat{\mathrm{Y}}_{\mathrm{ij}}\right).$$

To distinguish between the four possible lognormal-logistic (LL) estimators, subscripts based on the hierarchical components are used, as indicated in Table 2.

Table 2. The four lognormal-logistic small area estimators obtained by combining the dispersion parameter models

Logistic Lognormal	$\sigma_\omega = 0$	$\sigma_\omega > 0$
$\sigma_u = 0$	$\hat{\bar{Y}}_i^{LL_{00}}$	$\hat{\bar{Y}}_i^{LL_{0\omega}}$
$\sigma_u > 0$	$\hat{\bar{Y}}_i^{LL_{u0}}$	$\hat{\bar{Y}}_i^{LL_{u\omega}}$

Letting $\hat{T}$ denote the population total estimator of Karlberg (2000b), we have that

$$\hat{\mathrm{T}} = \sum_{\mathrm{i}=1}^{\mathrm{a}} \mathrm{N_i}\widehat{\overline{\mathrm{Y}}}_{\mathrm{i}}^{\mathrm{LL_{00}}}$$

where *a* is the number of areas. As the exact same model is used, the variance estimator of Karlberg (2000b) is easily applicable to $\widehat{\overline{\mathrm{Y}}}_{\mathrm{i}}^{\mathrm{LL_{00}}}$.

3. Empirical evaluation of estimator properties

3.1. Estimators evaluated and benchmark estimators

The lognormal-logistic estimators of small area means have been evaluated against estimators based on the raw (unlogarithmed) Y_{ij} values. For real survey data, we used

(i) the direct estimator $\hat{\bar{Y}}_i^{DIR}$, as implemented in the SAE package (Molina and Marhuenda, 2013)

(ii) the synthetic unit-level regression estimator $\hat{\bar{Y}}_i^{REG}$ (thus without area effect), used for benchmarking purposes by Karlberg (2000b) and

(iii) the Battese, Harter, Fuller estimator (1988) $\hat{\bar{Y}}_i^{BHF}$ as implemented in the said SAE package.

For random data, we limited the set of benchmark estimators to (ii) and (iii), since there was no model misspecification for the lognormal-logistic estimators rendering the direct estimator superior in terms of unbiasedness. Since there are two sets of auxiliary information **Ξ** and **X** used by the lognormal-logistic estimators, we used the union of those matrices as auxiliary information for the benchmark estimators (ii) and (iii) using auxiliary information.

3.2. Stochastic data

3.2.1. Lognormal-logistic parameters

There are numerous ways to vary the ways in which stochastic data are generated. In this simulation study, we fixed most parameters, in essence only varying the small area sample size n_i and, directly or indirectly, the dispersion parameters of the two types of area-level effects (u_i and ω_i).

First, we limited the study to lognormal-logistic data, saving the investigation of possible model misspecification to the simulation study related to real survey data. In terms of size, we used only a=20 small areas, and fixed the ratio between small area (population) size and small area sample size to N_i/n_i=20, and also imposed the restriction that n_i be the same across all of the a areas. Considering the essence of auxiliary variables being sufficiently captured by one auxiliary variable for the purposes of this simulation study, we limited the **Ξ** and **X** matrices to contain (in addition to the requisite intercept dummies) a sole auxiliary variable each. We set these variables to be i.i.d. normal distributed, i.e. $\Xi_{1ij} \sim N(0,1)$ and $X_{1ij} \sim N(0,1)$ (thus having zero correlation between the two auxiliary variables; $\rho_{X\Xi} = 0$).

We invariably used the logistic regression parameter **β**=(1,1); with the logistic intercept parameter β_0 thus equal to 1, the resulting number of non-zero Y_{ij} values is roughly equal to e/(1+e) ≈ ¾. We thus have roughly ¼ zero-valued observations in the population. We used the lognormal regression parameter **B**=(0,1) throughout.

3.2.2. Simulation study

With most parameters fixed, we tried out the Cartesian product of the following free parameters:

- We used two different area sample sizes n_i=20 and n_i=5.
- With the overall variance in the lognormal component fixed at

  $$\sigma_{\bullet}^2 = \sigma_u^2 + \sigma_e^2 = 1,$$

  we varied the area effect proportion

  $$p_\sigma = \sigma_u^2 / \sigma_{\bullet}^2$$

  in small increments from 0 to 0.2.

- We varied the logistic area effect standard deviation σ_ω in small increments from 0 to 1.5.

For each parameter combination, we generated K=100 random populations and drew a single stratified random sample from each of them. (However, if any sample with no positive observations at all for an entire area, i.e. where any n_{+i}=0, was encountered, the population was regenerated, and the sample was redrawn.) The three benchmark and four lognormal-logistic estimators were then used to estimate the small area averages, and for each area *i* and replicate *k*, the relative bias of the estimator EST was calculated as

$$\mathrm{RB}_{\mathrm{i(k)}}^{\mathrm{EST}} = \left(\hat{\bar{\mathrm{Y}}}_{\mathrm{i(k)}}^{\mathrm{EST}} - \bar{\mathrm{Y}}_{\mathrm{i(k)}}\right)/\bar{\mathrm{Y}}_{\mathrm{i(k)}}$$

and the relative MSE of EST was obtained as

$$\mathrm{RMSE}_{\mathrm{i(k)}}^{\mathrm{EST}} = \left(\mathrm{RB}_{\mathrm{i(k)}}^{\mathrm{EST}}\right)^2.$$

Thereafter, in view of the fact that with the stochastic data, the small areas are interchangeable, the overall relative bias of the estimator EST is obtained by averaging $RB_{i(k)}$ across all areas as well as across all replicates as:

$$\mathrm{RB}_{\mathrm{EST}} = \frac{1}{\mathrm{aK}}\sum_{\mathrm{k=1}}^{\mathrm{K}}\sum_{\mathrm{i=1}}^{\mathrm{a}} \mathrm{RB}_{\mathrm{i(k)}}^{\mathrm{EST}}$$

and the overall relative root mean squared error is obtained as:

$$\mathrm{RRMSE}_{\mathrm{EST}} = \sqrt{\frac{1}{\mathrm{aK}}\sum_{\mathrm{k=1}}^{\mathrm{K}}\sum_{\mathrm{i=1}}^{\mathrm{a}} \mathrm{RMSE}_{\mathrm{i(k)}}^{\mathrm{EST}}}\ .$$

The relative efficiency of an estimator EST w.r.t. a benchmark estimator BNCH, can then be obtained as

$$\mathrm{RE}_{\mathrm{BNCH}}^{\mathrm{EST}} = \mathrm{RRMSE}_{\mathrm{BNCH}}^2/\mathrm{RRMSE}_{\mathrm{EST}}^2\ .$$

3.2.3. Results

In Figure 1, the observed relative efficiency at an area level sample size n_i=20 for each dispersion parameter combination is illustrated for each estimator/benchmark estimator (columns; orange labels / rows; green labels) pair. In essence, green colour coding indicates superiority w.r.t. the benchmark, and red-orange-yellow patterns indicate various degrees of inferiority. Given the multitude of comparisons that we perform below, we will, for compactness, use the index as a shorthand form to refer to an estimator in running text; for instance, we let LL_{00} denote the estimator

$$\hat{\bar{\mathrm{Y}}}_{\mathrm{i}}^{\mathrm{LL_{00}}}$$

This largely corresponds to the row and column labels of the figures presenting the results (although the figures use “w” for ω, and have a leading “Y” for the estimators based on unlogarithmed values).

A reasonable conjecture is that there is monotonicity of the true relative efficiency w.r.t. to the dispersion parameters, meaning that if the number of replicate populations was larger, the colour regions would be contiguous. Match-ups where a colour mosaic is displayed are thus an indication of lack of precision in terms of RE estimation. The prevalence of such “mosaics” in Figure 1 thus means that we can only express ourselves in terms of general tendencies regarding the impact of dispersion parameters on the RE of an estimator w.r.t. another estimator. We would have to conduct a simulation study with somewhat more replicates to be able to more precisely define the boundaries at which one estimator becomes more efficient than the benchmark estimator.

However, already the general tendencies observed are quite informative. Starting out with the intra-class comparison among the lognormal-logistic estimators, we see, as expected, that if the logistic area dispersion parameter σ_ω increases (rightwards in each pane), the estimators incorporating σ_ω ($LL_{u\omega}$ and $LL_{0\omega}$) fare better than the corresponding estimators lacking those components (LL_{u0} and LL_{00}, respectively). The pairwise comparisons in question ($LL_{u\omega}$ vs. LL_{u0}; $LL_{0\omega}$ vs. LL_{00}) indicate that this superiority holds already for very small positive values of σ_ω, with the boundary somewhere around σ_ω=0.2. Similarly, an increase in the lognormal area effect proportion (upwards in each pane) renders the estimators incorporating a positive parameter σ_u ($LL_{u\omega}$ and LL_{u0}) more efficient than those that do not ($LL_{0\omega}$ and LL_{00}, respectively). The pairwise comparisons in question ($LL_{u\omega}$ vs. $LL_{0\omega}$; LL_{u0} vs. LL_{00}) indicate that this superiority occurs already at a very modest area effect proportion (the boundary seemingly falling somewhere around p_σ=0.025).

Turning our attention to comparisons with the design-unbiased (*DIR*) and model-based (*REG* and *BHF*) estimators based on raw, untransformed Y_{ij} values, it appears from Figure 1 that the lognormal-logistic estimator incorporating both variants of area-level effects, $LL_{u\omega}$, is more efficient than the estimators based on untransformed data, with the possible exception of situations where both σ_ω and σ_u are very small.

While Figure 1 presents the bottom line, i.e. the relative efficiency, it could also be interesting to explore the relative bias of the various estimators. The results (not shown here) indicate that, as expected, the relative bias of the direct estimator is invariably low regardless of the parameterisation – typically in the range of ±1%. At p_σ=0, as σ_ω increases from 0 to 1.5 the relative bias of the appropriate estimator $LL_{0\omega}$ increases only moderately (from 2% to 6%), whereas the bias of the estimator LL_{00}, which lacks a logistic area component, increases dramatically (from 2% to 30%). At p_σ=0.2 and σ_ω=0, the estimators lacking a lognormal area component have a relative bias of 20%, compared to a modest relative bias of 5% for those that allow for a positive value of σ_u .

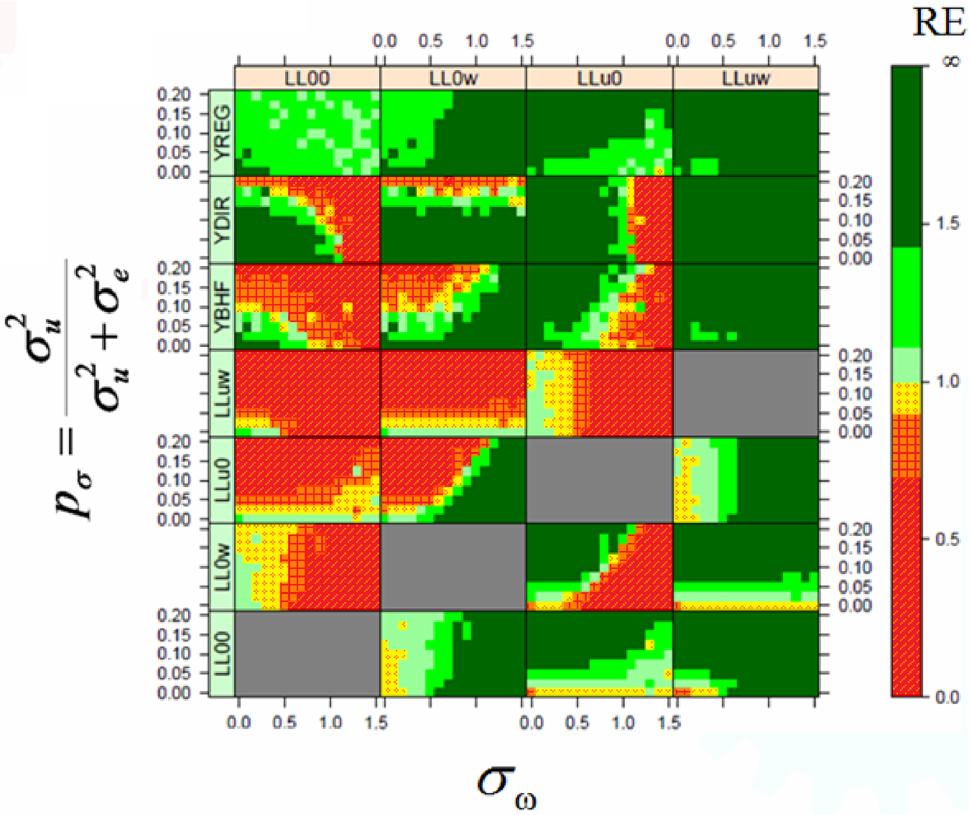

Figure 1. Relative Efficiency (*RE*) of each of the four evaluated estimators (columns) against seven benchmark estimators (rows) for various values of the lognormal-logistic parameters σ_ω and p_σ . Each rectangle corresponds to 100 stratified random samples; each of them drawn from a different lognormal-logistic data set. For each of the *a*=20 small areas (each with a size N_i=400), the sample size is n_i=20.

Figure 2 summarises the relative estimator efficiencies for random data with area level sample size of n_i=5 (with the sampling proportion remaining the same, the area population size N_i is 5·20=100 here, whereas it was 20·20=400 for the results summarised in Figure 1 above). To summarise the results for that very small sample size, we could say that the same general tendencies hold, but with the area-level dispersion parameter boundaries shifted upwards (to $\sigma_\omega \approx 0.3$ and $p_\sigma \approx 0.075$). However, Figure 2 is much more of a “mosaic” nature. This is due to the far more volatile nature of both numerator and denominator (in turn due to the high volatility of the small area estimators caused by the very low sample sizes for the small areas). A surprising finding is, however, that for very large values of the logistic dispersion parameter ($\sigma_\omega \approx 1.5$) the direct estimator turns out to be superior to those based on lognormal-logistic models. This might be attributed to the very low number of non-zero observations used to estimate the lognormal distribution parameters and area effects.

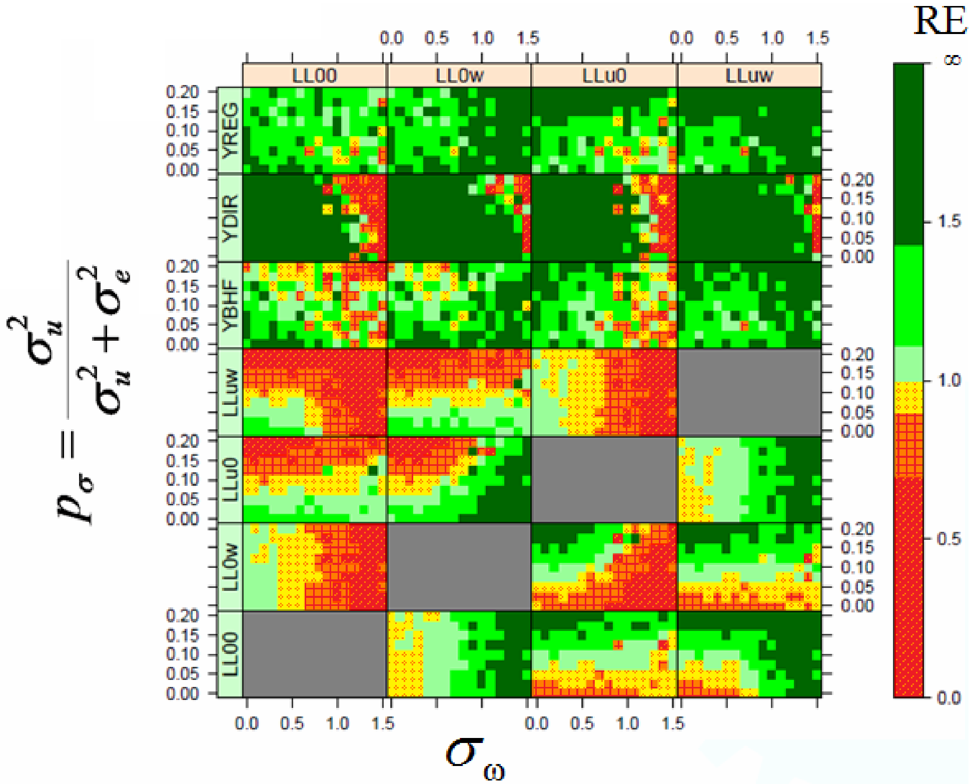

Figure 2. Relative Efficiency (*RE*) of each of the four evaluated estimators (columns) against seven benchmark estimators (rows) for various values of the lognormal-logistic parameters σ_ω and p_σ . Each rectangle corresponds to 100 stratified random samples; each of them drawn from a different lognormal-logistic data set. For each of the a=20 small areas (each with a size N_i=100), the sample size is n_i=5.

3.3. Survey data

3.3.1. The AAGIS data

Like, e.g. Chandra and Chambers (2005) and Chambers and Tzavidis (2006) and Molina (2009), we have applied our lognormal-logistic estimators data obtained from a sample of 1652 farms that participated in the Australian Agricultural and Grazing Industries Survey (AAGIS). This survey includes a number of variables with skewed distributions and a sizeable proportion of 0s, lending itself well to lognormal-logistic modelling. Moreover, as the data are subdivided into 29 regions (areas), it is also useful for Small Area Estimation. Out of the 1652 observations, we have excluded one with a zero-valued observation for a possible auxiliary variable (to allow us to logarithm it if needed). Some basic characteristics of the variable Beef Cattle are provided in Appendix 1.

The only possible *Y* variable for our class of estimators is Beef Cattle, since the other variables with zero-valued observations have some areas for which there are no observations with positive values at all, rendering estimation with the current implementation of the BHF estimator in the SAE package impossible. (Obviously, this would have to be resolved before such lognormal-logistic estimators are implemented in production.) We have used Farm Area as the auxiliary variable for the logistic component as well as for the lognormal one.

In the simulation study, we have drawn stratified samples (treating the AAGIS data, albeit they are from a sample survey, as a population of size 1651). The only parameter varied has been n_i , for which we have used six different parameterisations, of two different types: (i) the same absolute number across areas (capped at a sample fraction of 50% per area) and (ii) a constant sample fraction per area (with a minimum absolute sample size of 1).

For each parameterisation, we have used 100 replicates. It should be underlined that in contrast to the evaluation of estimator performance for random data (where the areas could be considered interchangeable), the performance measures have been calculated area by area (across all replicates), and not across all small areas. The area-specific relative bias of area *i* is thus obtained as

$$\mathrm{RB}_{\mathrm{EST;i}} = \frac{1}{\mathrm{K}}\sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{RB}_{\mathrm{i(k)}}^{\mathrm{EST}}$$

and the other performance measures are obtained analogously.

3.3.2. Results

As could be seen from Figure 3, the bias is severe for LL_{00} and $LL_{u\omega}$ for certain small areas, with the relative bias sometimes extremely high. With *DIR* unbiased by design, this inevitably carries over into the direct estimator being superior in terms of relative efficiency for such areas, as illustrated by Figure 4. Taking area 1, the area with the smallest number of positive observations (N_{+1}=4) as an example, we have that the relative bias of LL_{00} is around 100, which, in spite of the high variance of *DIR*, carries over a relative efficiency of the direct estimator of approximately 10^4.

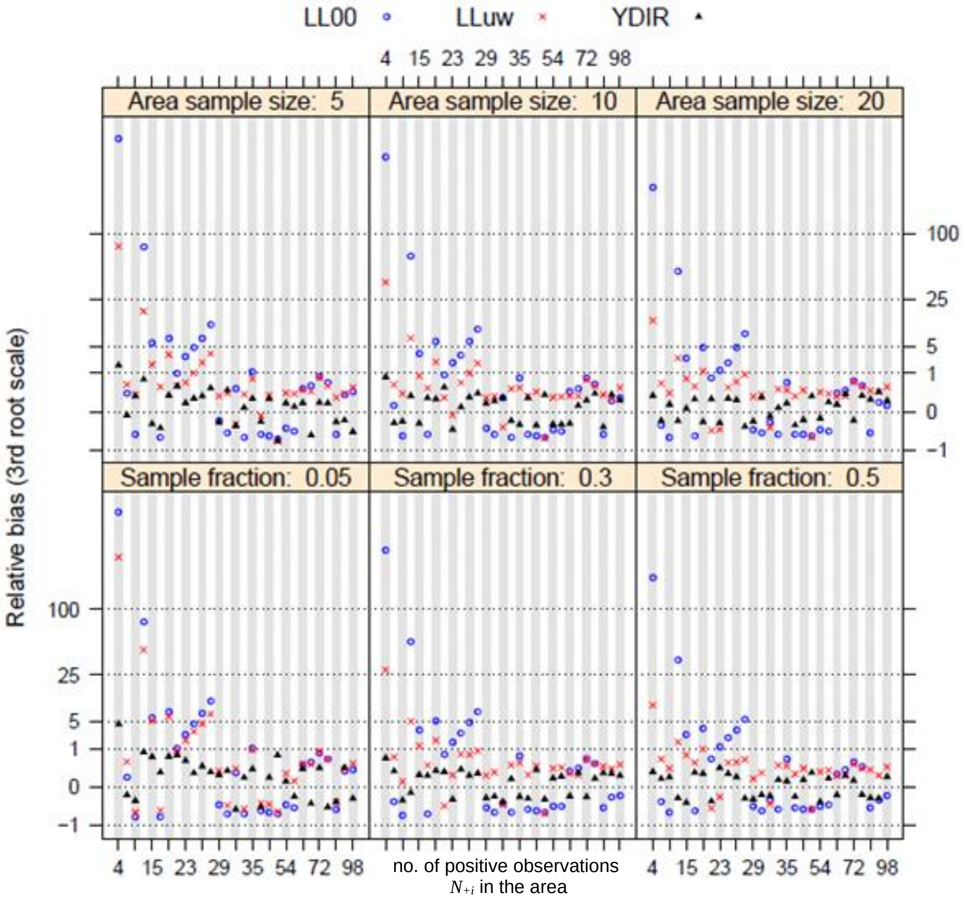

Figure 3. Average relative bias of the *DIR*, LL_{00} and $LL_{u\omega}$ estimators of Beef Cattle area means for various sample sizes. 100 replicates have been used for each sample size parameter.

Owing to these findings, we do not present findings regarding the other benchmark estimators or lognormal-logistic estimators here; if the lognormal-logistic estimators fail to outperform the direct estimators, their performance relative to each other and relative to other benchmark estimators becomes less interesting.

In Appendix 1, the drivers for these tendencies are investigated. In short, as is often the case for small area estimation (Chambers *et al.*, 2014), a model which works reasonably well at population level is found to be inappropriate at the area level.

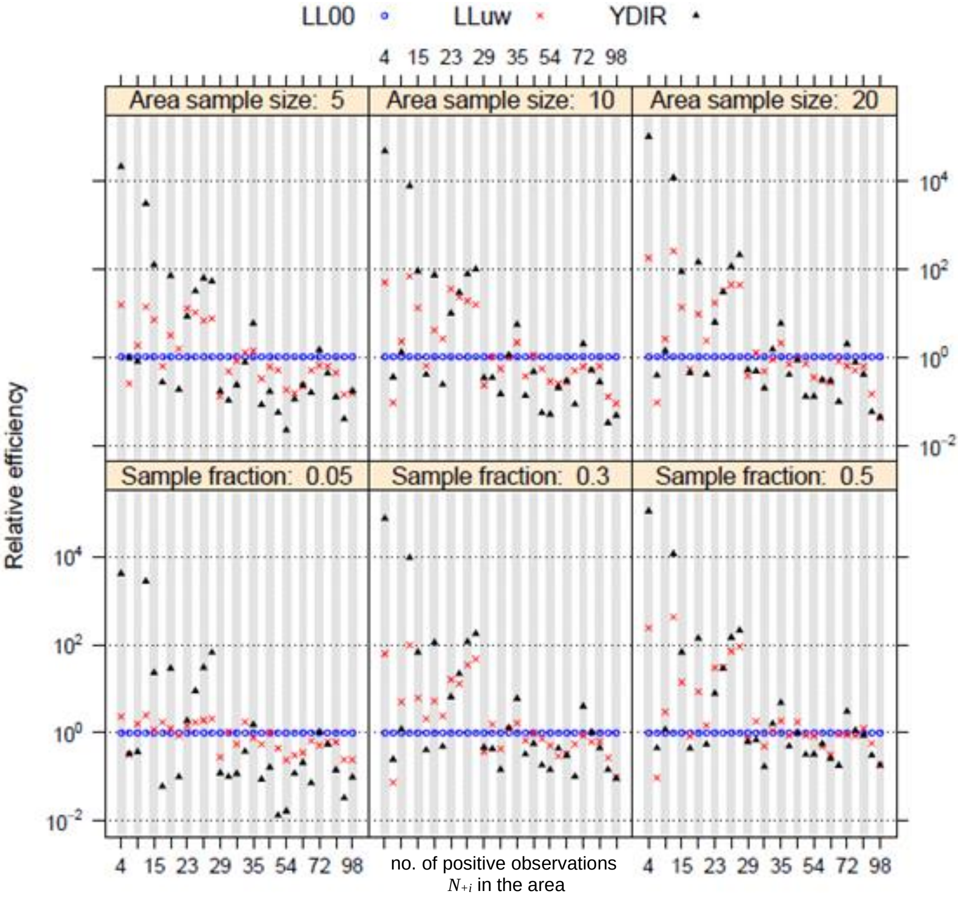

Figure 4. Relative efficiency of the *DIR* and $LL_{u\omega}$ estimators (w.r.t. LL_{00}) of Beef Cattle area means for various sample sizes. 100 replicates have been used for each sample size parameter

4. Conclusions

In Section 2 of this paper, we have arrived at four different lognormal-logistic estimators of small area means by combining the lognormal small area estimator of Berg and Chandra (2012) with the lognormal logistic model of Karlberg (2000b), and optionally incorporating hierarchical logistic regression.

We have conducted a simulation study to investigate the estimator properties under ideal circumstances, i.e. when the presumed lognormal-logistic model holds. As seen from Section 3.2, the estimators behave largely as predicted, i.e. when lognormal and/or logistic area-level effects are present, models incorporating such effects are superior, in terms of relative efficiency. Interestingly, this holds already for rather small effects; the “penalty” for “unnecessarily” estimating a parameter when such a parameter is not present (thus introducing “white noise” into the estimation process) seems to be very modest. Using $LL_{u\omega}$ for lognormal-logistic data thus seems to be the best option (with the possible exception of situations with very low sample sizes (say $n_i \leq 5$) combined with large heterogeneity between the areas in terms of the proportion of positive observations (say $\sigma_\omega \geq 1.5$) when the direct estimator might be a safer option).

However, the model assumptions could be challenged. First, the assumption about independence between the lognormal and logistic components, made in Section 2.1, could be challenged; Pfefferman *et al.* (2008) convincingly argue for assuming a correlation between the two types of random effects; an extension of the model presented in this paper following the Bayesian approach proposed by Pfefferman *et al.* to relax the independence assumption. Even more critical is the fact that in real life data do not necessarily comply with a lognormal-logistic model, rendering the possible presence of correlation an issue of secondary importance. As could be seen from Section 3.3, the estimator’s performance for the Beef Cattle variable of AAGIS is disastrous for certain small areas. This is studied in Appendix 1, where it is found that the small area estimation fails even if the model is fitted to the entire AAGIS data set, as going from national level to regional (area) level leads to severely biased estimates for some areas. Given this failure at small area population level, it is no surprise that the performance is bad when estimation is carried out for random samples. The situation is somewhat improved when area-level random effects are introduced – but an intolerable bias level remains for many areas.

It would be interesting to evaluate whether this is an artefact of the AAGIS data, i.e. if there are other real data sets where the lognormal-logistic estimators fare better, and what the properties of such data sets are (e.g. larger “small areas”, or more highly correlated variables) – or if this poor performance is all but unavoidable. It could be argued that the performance issues are not so much related to the data as to the model, and there are a number of possible improvements of the lognormal-logistic models, such as somehow integrating it into the robust weighted mixed model of Chandra and Chambers (2011), which might be worth exploring.

Minor possible improvements also include a more formal treatment of the bias correction factor (currently simply carried over from Berg and Chandra; 2012), and the development of a proper model-based variance estimator (currently only readily available for LL_{00}), possibly even with an uncertainty measure for this variance (see Royall and Cumberland 1978 and Fellner 1986). Practical extensions to allow for some n_{+i}=0, and extensions to also allow negative values of Y_{ij} are also worth considering.

REFERENCES

BARNETT, V., LEWIS, T., (1994). Outliers in Statistical Data, 3rd ed. John Wiley & Sons.

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, Vol. 83, pp. 28–36.

BERG, E., CHANDRA, H., (2012). Small area prediction for a unit level lognormal model, Federal Committee on Statistical Methodology Research Conference.

CARROLL, R., RUPPERT, D., (1988). Transformation and Weighting in Regression, Chapman and Hall.

CHAMBERS, R. L., (1986). Outlier robust finite population estimation, Journal of the American Statistical Association, Vol. 81, pp. 1063–1069.

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

CHAMBERS, R. L., DORFMAN, A. H., (2003). Transformed variables in survey sampling, Joint Statistical Meetings, Section on Survey Research Methods.

CHAMBERS, R. L., TZAVIDIS, N., (2006). M-quantile models for small area estimation., Biometrika, Vol. 93, pp. 255–268.

CHANDRA, H., CHAMBERS, R. L., (2005). Comparing EBLUP and C-EBLUP for Small Area Estimation, Statistics in Transition, Vol. 7, pp. 637–648.

CHANDRA, H., CHAMBERS, R. L., (2011). Small area estimation under transformation to linearity, Survey Methodology, Vol. 37, pp. 39–51.

CHEN, G., CHEN, J., (1996). A Transformation Method for Finite Population Sampling Calibrated with Empirical Likelihood, Survey Methodology, Vol. 22, pp. 139–146.

CHEN, J., CHEN, S.-Y., RAO, J. N. K., (2003). Empirical Likelihood Confidence Intervals for the Mean of a Population Containing Many Zero Values, The Canadian Journal of Statistics, Vol. 31, pp. 53–68.

FELLNER, W. H., (1986). Robust estimation of variance components, Technometrics, Vol. 28, pp. 51–60.

FULLER, W. A., (1991), Simple estimators for the mean of Skewed populations, Statistica Sinica, Vol. 1, pp. 137–158.

HIDIROGLOU, M. A., SMITH, P. A., (2005). Developing Small Area Estimates for Business Surveys at the ONS, Statistics in Transition, Vol. 7, pp. 527-539.

HIDIROGLOU, M. A., SRINATH, K. P., (1981). Some estimators of a population total from simple random samples containing large units, Journal of the American Statistical Association Vol. 76, pp. 690-695.

HUBER, P. J., (1981). Robust Statistics, John Wiley.

HUBERT, M., VAN DER VEEKEN, S., (2007). Outlier detection for skewed data, Journal of Chemometrics Vol. 22, pp. 235–246.

HUGGETT, M., (1996). Wealth distribution in life-cycle economies, Journal of Monetary Economics, Vol. 38, pp. 469–494.

JIANG, J., LAHIRI, P., (2006). Estimation of Finite Population Domain Means: A Model-Assisted Empirical Best Prediction Approach, Journal of the American Statistical Association, Vol. 101, pp. 301–311.

KARLBERG, F., (2000a). Population Total Prediction Under a Lognormal Superpopulation Model, Metron, Vol. LVIII, pp. 53–80.

KARLBERG, F., (2000b). Survey Estimation for Highly Skewed Populations in the Presence of Zeroes, Journal of Official Statistics, Vol. 16, pp. 229–241.

KOKIC, P. N., (1998). On Winsorisation in Business Surveys, SSC Annual Meeting, Proceedings of the Survey Methods Section.

LAMBERTA, D., (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, Technometrics, 34, pp. 1–14.

LEE, H. L., (1995). Outliers in Business Surveys, In Business Surveys Methods, edited by Cox, Binder, Chinnappa, Christianson, Colledege and Kott, Chapter 26. John Wiley.

LEHTONEN, R., SÄRNDAL C. E., VEIJANEN, A., (2003). The effect of model choice in estimation for domains, including small domains, Survey Methodology, Vol. 29, pp. 33–44.

MINCER, J., (1970). The Distribution of Labor Incomes: A Survey With Special Reference to the Human Capital Approach, Journal of Economic Literature 8, pp. 1–26.

MOLINA, I., (2009). Uncertainty under a multivariate nested-error regression model with logarithmic transformation, Journal of Multivariate Analysis, Vol. 100, pp. 963–980.

MOLINA, I., Marhuenda, Y., (2013). Package ‘sae’, http://cran.r project.org/web/packages/sae/sae.pdf

PFEFFERMANN, D., (2013). New Important Developments in Small Area Estimation, Statistical Science 28, pp. 40–68.

PFEFFERMANN, D., Terryn, B. Moura, F. A. S., (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries, Survey Methodology, Vol. 34, pp. 235–249.

RAO, J. N. K., (2003), Small Area Estimation, Wiley.

ROYALL, R. M., (1982), Finite populations (Sampling from), Entry in the Encyclopedia of Statistical Sciences.

ROYALL, R. M., CUMBERLAND, W. G., (1978). Variance estimation in finite population Sampling, Journal of the American Statistical Association Vol. 71, pp. 351–358.

RÖNNEGÅRD, L., SHEN, X. ALAM, M., (2010). hglm: A Package for Fitting Hierarchical Generalized Linear Models, The R Journal Vol. 2, pp. 20-28, http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Roennegaard~et~al.

SALVATI, N., Chandra, H., Chambers, R. L., (2012). Model Based Direct Estimation of Small Area Distributions, Australian & New Zealand Journal of Statistics 54, pp. 103–123.

SEARLS, D. T., (1966). An estimator which reduces large true observations, Journal of American Statistical Association, Vol. 61, pp. 1200–1204.

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

SHLOMO, N., PRIAM, R., (2013). Improving Estimation in Business Surveys. Chapter 4.2, 52–70 in BLUE-ETS Deliverable D6.2: Best practice recommendations on variance estimation and small area estimation in business surveys, edited by R. Bernardini Papalia, C. Bruch, T. Enderle, S. Falorsi, A. Fasulo, E. Fernandez-Vazquez, M. Ferrante, , J.P. Kolb, R. Münnich, S. Pacei, R. Priam, P. Righi, T. Schmid, N. Shlomo, F. Volk and T. Zimmermann.

SLUD, E., MAITI, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 68, pp. 239–257.

THORBURN, D., (1993). The treatment of outliers in economic statistics, Proceedings of the International Conference on Establishment Surveys, Buffalo, New York.

WANG, J., FULLER W. A., (2003). The Mean Squared Error of Small Area Predictors Constructed with Estimated Area Variances." Journal of the American Statistical Association, Vol. 98, pp. 716–723.

YOUNG, K. H., YOUNG, L. Y., (1975). Estimation of Regressions Involving Logarithmic Transformation of Zero Values in the Dependent Variable, The American Statistician , Vol. 29, pp. 118–120.

ZIMMERMANN, T., Münnich, R., (2013). Coherent small area estimates for skewed business data, Proceedings of the 2013 European Establishment Statistics Workshop.

APPENDIX

Appendix 1. Methodological details

A.1. Regression line fit at population level

In an attempt to identify the root cause of the poor performance of the lognormal-logistic estimators, we started out by fitting the model associated with LL_{00} to the entire population, i.e. the 1651 AAGIS observations. As could be seen from Figure A1, the model fits the data reasonably well; this is corroborated by the performance of $\hat{T}$ for the very same variables observed by Karlberg (2000b).

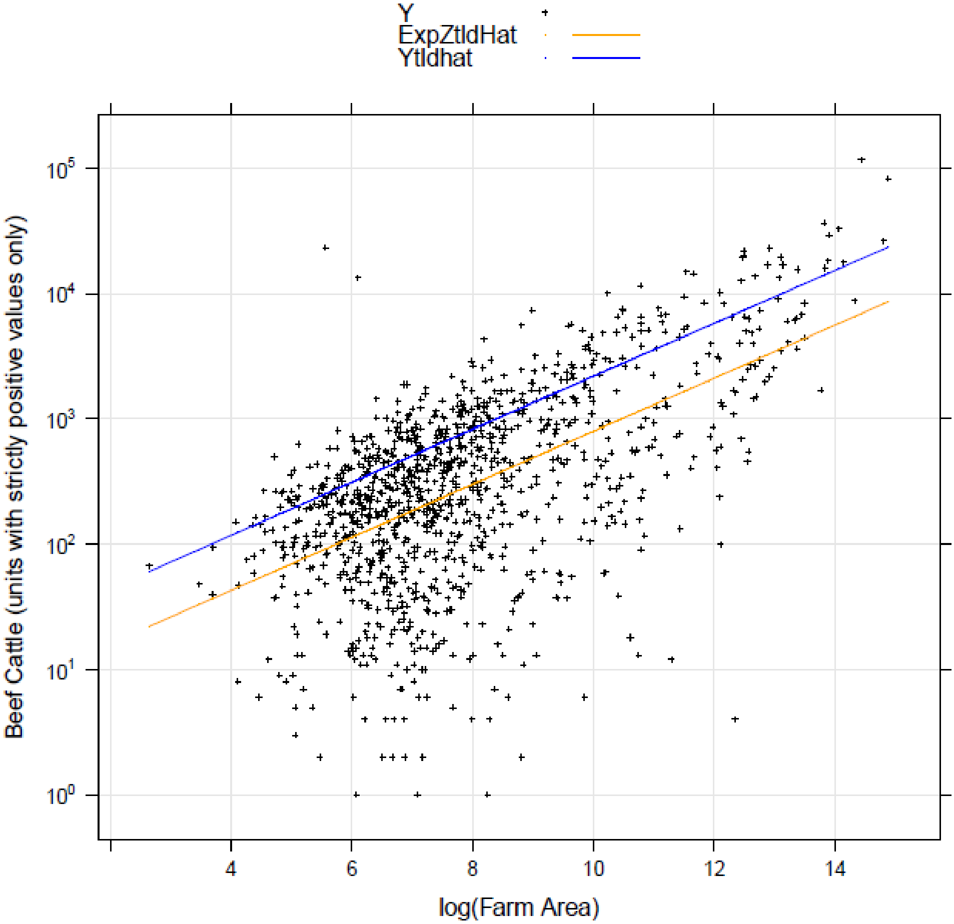

Figure A1. Regression line (red) fit to the logarithmed positive values of Beef Cattle for all 1651 observations (black) of AAGIS. The application of the bias correction factor is illustrated by the blue line.

A.2. Estimator performance at area level

Figure A2 demonstrates the effect of proceeding to the area level. There, we see that sometimes (taking the 4/43 area with 4 positive and 43 zero observations at the bottom left as an example) the entire area is composed of observations far from the regression line. Further investigations (not explicitly presented here) demonstrate that even if the large heterogeneity between areas in terms of zero valued observations (ranging from 0% to 91%, as could be seen from Table A1) is disregarded, the model completely fails to capture the structure of the positive values in a number of areas.

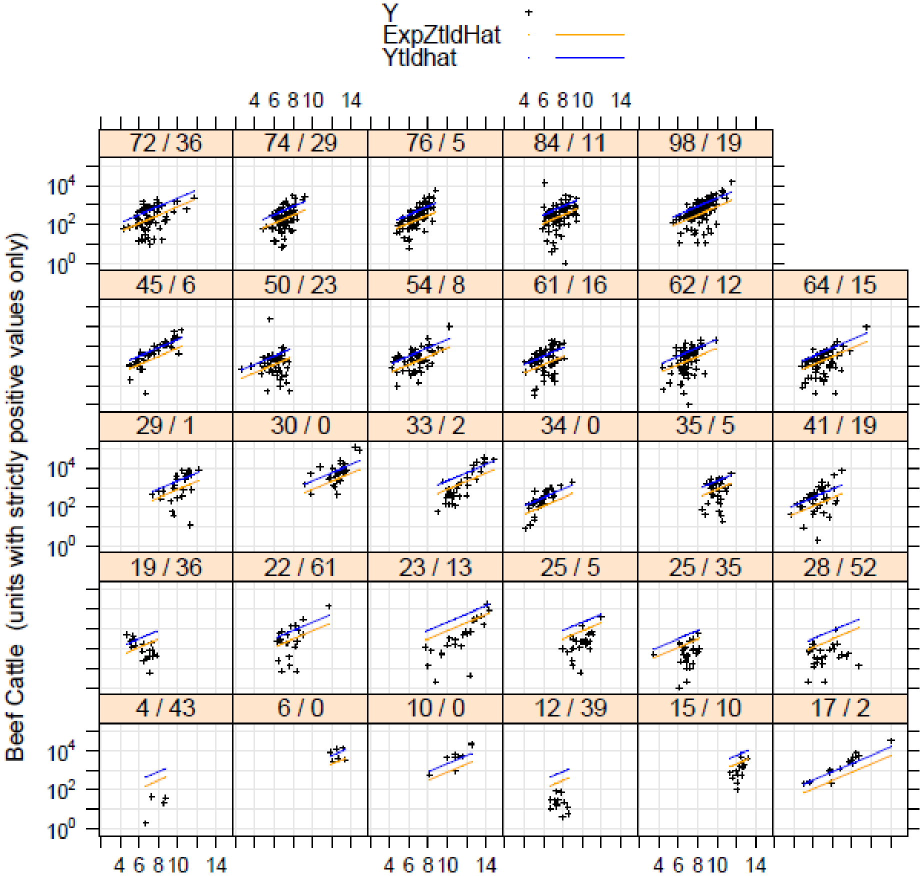

Figure A2. Regression line (red) fit to the logarithmed positive values of Beef Cattle for all 1651 observations of AAGIS, illustrated together with the observations (black) area by area. The number of positive/zero observation per area is indicated in the red strip above each area.

Obviously, if there is a severe bias even in an ideal situation, even with the model fit to the entire population, this is what could be expected to hold on average for samples drawn from that population as well. This is precisely what we observe in Figure 3 for certain of the areas in the simulation study.

As the incorporation of area effects allows the fitting of a model that is closer to the values observed for each area, the bias of LL_{00} is, as could be seen from Figure 3, somewhat less severe across most areas, in particular the smaller ones. However, the performance is still unacceptable for that estimator as well.

Table A1. Some characteristics of the AAGIS variable Beef Cattle

Area i	No. of farms N_i	$(N_i - N_{+i})/N_i$	$\sum_{\mathrm{j=1}}^{\mathrm{N}_i} \mathrm{Y_i}/N_{+i}$
1	47	91%	26.5
2	6	0%	7523.5
3	10	0%	8945.7
4	51	76%	28.8
5	25	40%	1554.7
6	19	11%	4285.6
7	55	65%	136.6
8	83	73%	1148.9
9	36	36%	1985.5
10	30	17%	430.1
11	60	58%	100.2
12	80	65%	97.8
13	30	3%	2774.7
14	30	0%	12903.0
15	35	6%	5878.8
16	34	0%	404.5
17	40	13%	1129.4
18	60	32%	670.5
19	51	12%	1139.6
20	73	32%	643.6
21	62	13%	530.9
22	77	21%	387.0
23	74	16%	390.7
24	79	19%	434.8
25	108	33%	435.2
26	103	28%	415.5
27	81	6%	526.6
28	95	12%	632.5
29	117	16%	980.6
All areas	**1651**	**30%**	**1308.5**

STATISTICS IN TRANSITION new series and SURVEY METHODOLOGY
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 563–584

BORROWING INFORMATION OVER TIME IN BINOMIAL/LOGIT NORMAL MODELS FOR SMALL AREA ESTIMATION

Carolina Franco[1], **William R. Bell**[2]

ABSTRACT

Linear area level models for small area estimation, such as the Fay-Herriot model, face challenges when applied to discrete survey data. Such data commonly arise as direct survey estimates of the number of persons possessing some characteristic, such as the number of persons in poverty. For such applications, we examine a binomial/logit normal (BLN) model that assumes a binomial distribution for rescaled survey estimates and a normal distribution with a linear regression mean function for logits of the true proportions. Effective sample sizes are defined so variances given the true proportions equal corresponding sampling variances of the direct survey estimates. We extend the BLN model to bivariate and time series (first order autoregressive) versions to permit borrowing information from past survey estimates, then apply these models to data used by the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program to predict county poverty for school-age children. We compare prediction results from the alternative models to see how much the bivariate and time series models reduce prediction error variances from those of the univariate BLN model. Standard conditional variance calculations for corresponding linear Gaussian models that suggest how much variance reduction will be achieved from borrowing information over time with linear models agree generally with the BLN empirical results.

Key words: area level model, complex surveys, American Community Survey, bivariate model, SAIPE.

1. Introduction

Small area estimation by area level models often uses linear Gaussian mixed models, specifically the model of Fay and Herriot (1979). When such models are applied to data from a repeated survey the question arises as to whether better results may be obtained by borrowing information from past data. Time series extensions to the Fay-Herriot (FH) model have thus been explored. See, e.g., Ghosh et al. (1996), Datta, Lahiri, Maiti, and Lu (1999), Saei and Chambers (2003), Rao

[1] U.S. Census Bureau. E-mail: Carolina.Franco@census.gov.
[2] U.S. Census Bureau. E-mail: William.R.Bell@census.gov.

and Molina (2015, Sections 4.4.3, 8.3, and 10.9), Esteban et al. (2012), and Pratesi et al. (2010, Chapter 3). Huang and Bell (2012) investigated the use of a bivariate FH model that, for each area, borrowed information from an estimate obtained by pooling recent past survey samples, which is similar to borrowing information from an average of past survey estimates.

Area level modeling has also been extended through the use of Generalized Linear Mixed Models (GLMM), which have been discussed in the context of small area estimation by Ghosh, et al. (1998) and Rao and Molina (2015, Section 10.13). GLMMs have potential advantages for modeling inherently discrete data arising from direct survey estimates of the number of persons that possess a certain characteristic (e.g., the number of persons in poverty). This can also be thought of as modeling survey estimates of the corresponding proportions (e.g., poverty rates). Directly applying a linear Gaussian model to such data may risk producing nonsensical results such as negative predictions or, more likely, prediction intervals that include negative values. Taking logarithms can eliminate these problems but creates the problem of dealing with direct estimates of zero that arise when no one in an area's sample possesses the characteristic whose prevalence is being estimated. Analogous problems arise if predicted proportions or their interval limits exceed one, or if direct estimates of proportions equal one. GLMMs avoid such problems and may also help account for the skewness typically inherent in such data when the underlying proportion is near zero or one.

This paper focuses on small area models that combine both extensions just mentioned. To address the challenges posed by discrete survey data, we use a binomial/logit normal (BLN) model. This particular GLMM assumes a binomial distribution for discrete observations, and a normal distribution with a linear regression mean function for logits of the binomial proportions. We determine effective sample sizes for the binomial distributions to preserve sampling variances estimated via a generalized variance function. To borrow information from past data we extend the BLN model to a bivariate version and then to a time series version. The latter uses a first order autoregressive model (AR(1)), although other time series structures could be used. The normality assumption for the random effects in the logits of the proportions facilitates these extensions for modeling dependence. One qualification to note is that the extensions assume independence of the sampling errors of the survey estimates for all years covered by the time series model, as well as for the two equations of the bivariate model.

Our motivating application comes from the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program. SAIPE provides annual poverty estimates for various age groups for states, counties, and school districts of the U.S.

An important SAIPE product is school district age 5–17 poverty estimates used by the U.S. Department of Education in allocating federal funds (over $14 billion in 2013) to school districts. For more information on the SAIPE program, see Bell et al. (2015) or the SAIPE web page at http://www.census.gov/did/www/saipe/.

The survey data source used by SAIPE, which we also use here for illustration, is poverty estimates from the American Community Survey (ACS). The ACS is the largest household sample survey in the United States, sampling approximately 3.5 million addresses per year. It collects data on a broad range of population characteristics such as income, health insurance coverage, and education, and publishes estimates annually. For areas with populations of $65,000$ or more, ACS publishes estimates based on a single year of data collection. For the smallest places, published estimates use data pooled from five years of ACS samples. The ACS, with its 5-year estimates, has effectively replaced the decennial census long form sample, which was last carried out in Census 2000. SAIPE poverty models use ACS 1-year estimates, which are not publicly released for counties with populations less than 65,000.

We focus here on modeling county poverty for school-aged (5-17) children, a key component of developing the SAIPE poverty estimates for school districts. The SAIPE production model has as a covariate the log of the Census 2000 long form county estimates of age 5–17 children in poverty. This covariate is going further and further out of date, motivating consideration of replacing it with past, but more recent, ACS data. Huang and Bell (2012) thus explored bivariate FH models for current ACS 1-year and past ACS 5-year poverty estimates. This issue also motivates the bivariate and time series extensions to the BLN model that we study here.

Our interest in studying the BLN model applied to SAIPE data stems from its potential advantages for modeling discrete data discussed above, a relevant consideration for the ACS 1-year estimates for small counties. Slud (2000, 2004) did several analyses comparing results from GLMM models to results from models similar to the SAIPE county production model. Slud (2000) showed advantages to the use of a unit level BLN model of sampled counts compared to a linear Fay–Herriot model for logged counts when the data were simulated from the GLMM model.

The rest of the paper proceeds as follows. Section 2 presents the BLN model and its extensions to bivariate and time series (AR(1)) versions. Section 3 presents results from application of these models to ACS county poverty data for 2012. We compare results between the variants of the BLN model to illustrate the potential benefits of the two different ways of borrowing information from past data. Section 4 provides conclusions.

2. Binomial/Logit Normal (BLN) Models

The BLN model may be written as

$$\begin{aligned} y_i|p_i,n_i &\sim \text{Bin}(n_i,p_i) \qquad i=1,\ldots,m & (1)\\ \text{logit}(p_i) &= \mathbf{x}_i'\beta + u_i & (2) \end{aligned}$$

where $\text{logit}(p_i) = \log[p_i/(1-p_i)]$, $u_i \sim i.i.d.\ N(0,\sigma_u^2)$, and n_i is the sample size for area i. The model as given by (1)–(2) can be readily applied to unweighted sample counts y_i, but doing this ignores any complex aspects of the survey design. For applications to complex survey data where the y_i are survey weighted estimates, two problems arise. First, the possible values for the y_i will not be the integers $0,1,\ldots,n_i$ for any direct definition of sample size n_i. Instead, y_i will take a value from a finite set of unequally-spaced numbers (not necessarily integers) determined by the survey weights that apply to the sample cases in area i. Second, the sampling variance of y_i implied by the binomial distribution in (1), $n_i p_i(1-p_i)$, will be incorrect.

To address these problems we start by defining an "effective sample size" $\tilde{n}_i$, and an "effective sample number of successes" $\tilde{y}_i$, determined to maintain: (*i*) the direct survey weighted estimate $\tilde{p}_i$ of the true proportion, and (*ii*) a corresponding sampling variance estimate, $\widehat{\text{var}}(\tilde{p}_i)$. For the latter we set

$$\tilde{n}_i = \breve{p}_i(1-\breve{p}_i)\big/\widehat{\text{var}}(\tilde{p}_i) \qquad (3)$$

where $\breve{p}_i$ is a preliminary model-based prediction of the population proportion p_i (on which $\text{var}(\tilde{p}_i)$ truly depends), and $\widehat{\text{var}}(\tilde{p}_i)$ depends on $\breve{p}_i$ through a fitted generalized variance function (GVF). Franco and Bell (2013) give a detailed explanation of the implementation of this GVF for application of the BLN models to the ACS county poverty data used in SAIPE models. Liu, Lahiri, and Kalton (2007) and You (2008) used essentially this type of sampling variance model, but applied it in models of survey estimates of proportions assumed to follow either a normal or a Beta distribution.

Having thus determined $\tilde{n}_i$, we set $\tilde{y}_i = \tilde{n}_i \times \tilde{p}_i$ and, after rounding, substitute $(\tilde{n}_i,\tilde{y}_i)$ for (n_i,y_i) in (1). Note that $\tilde{y}_i = 0$ if $\tilde{p}_i = 0$, but this does not cause problems since the BLN allows for observations of zero. Moreover, $\breve{p}_i > 0$ in (3) implies $\tilde{n}_i > 0$ even if $\tilde{p}_i = 0$. Rounding of $\tilde{n}_i$ and $\tilde{y}_i$ may be required by computer software for the fitting of models such as (1)–(2).

We extend the univariate BLN given by (1)–(2) to a bivariate BLN, written as

$$\tilde{y}_{1i}|p_{1i},\tilde{n}_{1i} \sim \text{Bin}(\tilde{n}_{1i},p_{1i}) \qquad \tilde{y}_{2i}|p_{2i},\tilde{n}_{2i} \sim \text{Bin}(\tilde{n}_{2i},p_{2i}) \tag{4}$$

$$\text{logit}(p_{1i}) = \mathbf{x}'_{1i}\beta_1 + u_{1i} \qquad \text{logit}(p_{2i}) = \mathbf{x}'_{2i}\beta_2 + u_{2i} \tag{5}$$

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim i.i.d.\ N(0,\Sigma), \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

for $i = 1,\ldots,m$. In (4), for each area i $\tilde{n}_{1i}$ and $\tilde{y}_{1i}$ are the effective sample size and effective number of successes derived as discussed above from a direct survey estimate y_{1i} and a corresponding sampling variance estimate. Similarly, $\tilde{n}_{2i}$ and $\tilde{y}_{2i}$ are derived from another direct survey estimate y_{2i} and corresponding sampling variance estimate. The bivariate BLN model can be applied to estimates y_{1i} and y_{2i} from two different surveys or for two different time points from the same repeated survey. Notice, though, that $\tilde{y}_{1i}$ and $\tilde{y}_{2i}$ are assumed conditionally independent (given $p_{1i},\tilde{n}_{1i}$ and $p_{2i},\tilde{n}_{2i}$), as will be the case if the samples on which they are based are drawn independently. This is true for our application of the bivariate BLN in Section 3, where y_{1i} and y_{2i} are ACS 1-year and previous 5-year poverty estimates, respectively, since ACS samples are drawn approximately independently each year. (The ACS housing unit samples are drawn independently each year from one of five population subframes to which U.S. residential addresses are randomly assigned, with rotation of the subframes on a five-year cycle. Sampling fractions for most areas are 5% or less. See U.S. Census Bureau (2014, pp. 32–46).)

Instead of summarizing the information in five prior years of ACS data through the resulting 5-year estimates, a logical alternative to consider is to use the corresponding five individual 1-year estimates. Putting this together with the current 1-year estimates, implies modeling six years of ACS 1-year estimates. We do this by extending the BLN to assume the model errors u_{it} have an AR(1) correlation structure:

$$\tilde{y}_{it}|p_{it},\tilde{n}_i \sim \text{Bin}(\tilde{n}_{it},p_{it}) \qquad i = 1,\ldots,m, \quad t = 1,\ldots,T \tag{6}$$

$$\text{logit}(p_{it}) = \mathbf{x}'_{it}\beta_t + u_{it} = \mathbf{x}'_{it}\beta_t + \sigma_t\tilde{u}_{it} \tag{7}$$

$$\tilde{u}_{it} = \phi\tilde{u}_{i,t-1} + \varepsilon_{it} \tag{8}$$

where $-1 < \phi < 1$. The ε_{it} are assumed distributed as $i.i.d.\ N(0,1-\phi^2)$ so that $\text{var}(\tilde{u}_{it}) = 1$ (Box and Jenkins 1970, p. 58) and $\text{var}(u_{it}) = \sigma_t^2$. Note that this version of the BLN-AR(1) model has different regression coefficients (β_t) and different model variances (σ_t^2) each year. We have three reasons for making this assumption. First, the true regression coefficients and model variances may actually differ year-to-year. Second, this assumption is implicitly made in current SAIPE production by fitting the univariate production models separately for each year. Third, and most importantly here, the assumption facilitates comparisons of results, especially the comparisons of posterior variances and standard deviations that we make in Section 3, to corresponding results obtained from the univariate and bivariate BLN models. Both the univariate and bivariate BLN models use regression coefficients and a model variance specific to the prediction year.

A more conventional version of the BLN-AR(1) model would set $\beta_t = \beta$ and $\sigma_t^2 = \sigma_u^2$ for all years t in the model. With this assumption, the covariance matrix of $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})'$ has the general form (Box and Jenkins 1970, pp. 56-58)

$$\operatorname{var}(\mathbf{u}_i) = \sigma_u^2 \begin{bmatrix} 1 & \phi & \cdots & \phi^{T-1} \\ \phi & 1 & \cdots & \phi^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \cdots & 1 \end{bmatrix}. \tag{9}$$

For the heteroscedastic version given by (7)–(8), we drop σ_u^2 in (9) and pre- and post-multiply the matrix there by a diagonal matrix with diagonal elements σ_t. Linear Gaussian models with AR(1) model errors for ACS poverty data were investigated by Taciak and Basel (2012) for application to logs of ACS county 5-17 poverty estimates, and by Hawala and Lahiri (2012) for application to ACS estimates of county 5-17 poverty rates. Esteban et al. (2012) applied such models to data from the Spanish Living Conditions Survey to improve direct survey estimates of the male and female poverty rates for Spanish provinces.

3. Application: Borrowing Information from Past Data in Small Area Estimation of Poverty for U.S. Counties

To illustrate the potential for variance reductions from the bivariate and AR(1) extensions to the BLN models, we apply these models to estimating poverty rates for school aged children in U.S. counties in 2012. The univariate BLN (1)–(2) models the 2012 ACS 1-year county poverty estimates, the bivariate BLN (4)–(5) models these estimates together with the 2007–2011 ACS 5-year county poverty estimates, and the BLN-AR(1) (6)–(8) models the ACS 1-year county poverty estimates from 2007–2012. We shall compare prediction results from these models for 2012 for 3,136 counties, omitting 6 counties from the SAIPE universe which were not consistently defined across all 6 years of data. We did the same analysis with data corresponding to prediction years 2010 and 2011 and obtained very similar results.

The regression variables used in each of the models included 1 for an intercept term, and logistic transformations of the following:

- the proportion of child tax exemptions "in poverty" for the county, i.e., the ratio of the number of child exemptions claimed on tax returns whose adjusted gross income falls below the poverty threshold divided by the total number of child exemptions for the county. (Notes: (*i*) In general terms, a "child tax exemption" is a child listed on an income tax return who is economically dependent on the person filing the return. (*ii*) The poverty threshold used is that applicable to a family of the size implied by the number of exemptions (persons) listed on the tax return.)

- an adjusted version of the county "child tax filer rate," which is defined as the number of child exemptions in the county claimed on tax returns divided by the county population age 0–17.

- the "SNAP participation ratio," defined as the ratio of county recipients of benefits from the Supplemental Nutrition Assistance Program (SNAP), a program that subsidizes food expenses of low income persons, in July of the previous year to the county population of the previous year.

Huang and Bell (2012) used the above ratio variables in bivariate models for ACS poverty rates, while Bell et al. (2007) used their logarithms in models for logs of ACS poverty rates. For $\mathbf{x}_{2i}$ in equation (5) of the bivariate BLN, we used the above variables defined for the middle year (2009) of the 5-year interval.

An issue arises for the child tax filer rate in that it often exceeds 1 due to the number of child tax exemptions in a county exceeding the county's age 0–17 population. This occurs because the upper age limit for a child tax exemption can exceed 17, ranging as high as 23 for university students, and with no age limit for disabled children. The issue was addressed by multiplying all child tax filer rates by a constant factor to bring the maximum rescaled filer rate just below 1, permitting the logistic transformation. This adjustment is discussed further in Franco and Bell (2013).

We used the JAGS software (Plummer 2010) to implement the three models via a Bayesian approach with noninformative priors. Regression parameters were given normal priors with large variances, while the random effect variances in our models were given flat priors on intervals $[0,\kappa]$ chosen wide enough to contain essentially all the posterior probability as judged from examination of their posterior densities for a univariate model. The parameters ρ of the bivariate BLN and ϕ of the BLN-AR(1) models were given flat priors on $(-1,1)$. We determined the effective sample sizes $\tilde{n}_i$ and effective numbers of successes $\tilde{y}_i$ for the BLN models as discussed in Section 2 for both the ACS 1-year and ACS 5-year poverty estimates. We used separately fitted GVFs for the sampling variances for each year of the 1-year estimates, as well as a separately fitted GVF for the variances of the 5-year estimates.

3.1. Variance reductions from the extensions of the BLN model

Figure 1 compares the posterior means and standard deviations obtained from JAGS for the rates of school-aged children in poverty for U.S. counties in 2012 from the univariate, bivariate, and AR(1) BLN models. Parts (a) and (b) show that the posterior means are similar regardless of which of the models we choose. Figure 1(c) shows the posterior standard deviations tend to be lower for the bivariate BLN model than for the univariate BLN model, suggesting some value to incorporating the ACS 5-year estimates into the model. The gains are modest, however. The average percentage reduction in posterior standard deviations from using the bivariate versus the univariate model is approximately 5%, with about an 11% corresponding average reduction in posterior variances. The AR(1) model, on the other hand, yields only a 2.3% average decrease in standard deviations and a 4.6% decrease in variances compared to the univariate model. On average, it has larger posterior standard deviations than does the bivariate model, as reflected in Figure 1(d).

As the returns from using the bivariate or AR(1) BLN models to borrow information from past data are so modest, the question arises as to whether the data provide much evidence of dependence over time in the model errors u_{it}. In fact, the posterior mean of ρ from the bivariate BLN is .51 with a 95% posterior (credible) interval of $(.43, .60)$, while the posterior mean of ϕ from the AR(1) model is 0.44, with a 95% interval of $(.39, .50)$. So the data provide clear evidence of dependence over time in the u_{it}, but modeling this dependence does not produce much reduction in prediction uncertainty for the county 5–17 poverty rates.

3.2. How much improvement should we expect from borrowing information from past data?

As a rough guide to how much improvement might be expected from the bivariate or AR(1) models over the univariate model, we consider the linear FH model case when the true dependence structure is a stationary AR(1) model and all model parameters are known. We also assume for simplicity that the model error variance σ_u^2 and the sampling variances v_i remain constant over time. For this case it is straightforward to compute and compare the posterior variances (prediction MSEs) for the univariate, bivariate, and AR(1) versions of the FH model using standard results on

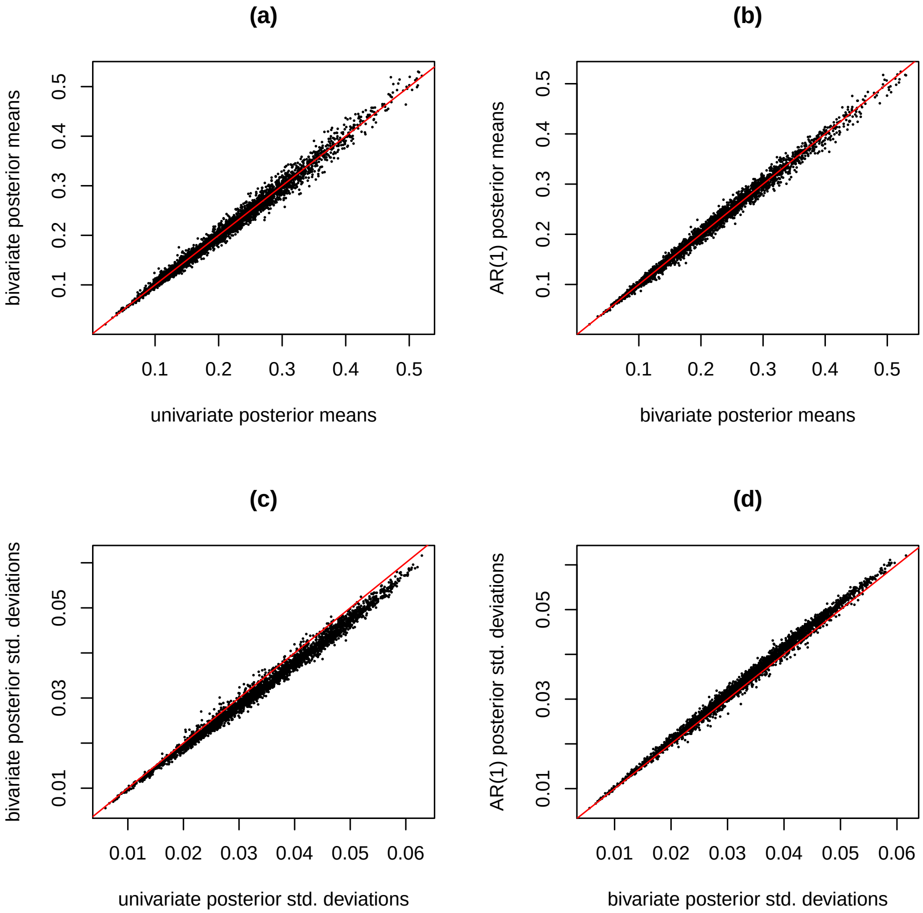

Figure 1: Comparison of posterior means and standard deviations for 2012 U.S. county poverty rates of school-aged children for univariate, bivariate, and AR(1) BLN Models.

conditional variances in a multivariate normal distribution – see the Appendix. Note that, since model parameters are assumed known, the predictions for each model are optimal conditional on the data used, but the data conditioned on differs across the three models.

Percent reductions in posterior variances for the bivariate and AR(1) models compared to the univariate model depend only on the parameter ϕ and variance ratio v_i/σ_u^2. Figure 2(a) shows contour plots of the percent variance reductions achieved by the AR(1) model as functions of ϕ and v_i/σ_u^2. (The plot assumes $\phi \geq .4$; a mirror image results for $\phi \leq -.4$, and percent reductions are small for $|\phi| < .4$.) It shows

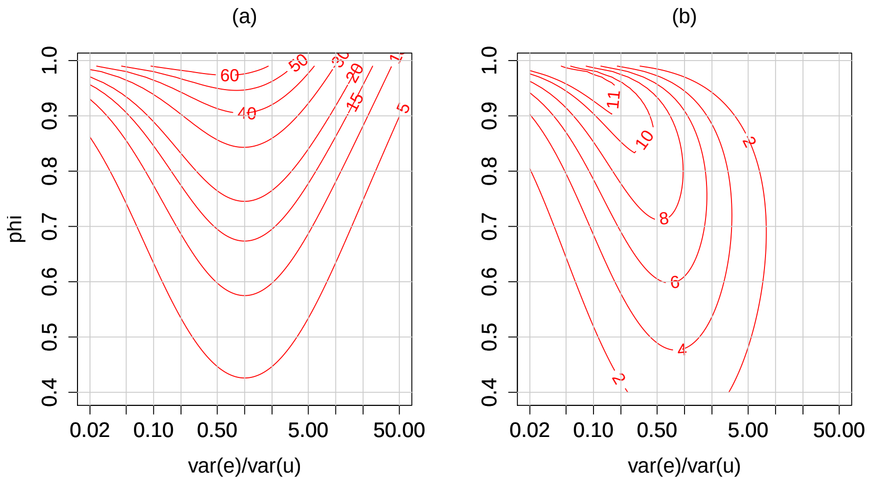

Figure 2: Contour plots of posterior variance percent reductions for small area estimates achieved by the FH-AR(1) model using 6 years of data compared to the univariate and bivariate FH models, when the true population characteristics actually follow an AR(1) model. Contours are shown as functions of the AR(1) parameter ϕ and $\mathrm{var}(e_i)/\mathrm{var}(u_i)$, the ratio of the sampling error variance to the model error variance. (a) Reductions from the AR(1) versus the univariate model. (b) Reductions from the AR(1) versus the bivariate model.

that the variance reductions increase with increasing values of ϕ, and decrease as the value of v_i/σ_u^2 deviates from 1.0. (Note that the x-axis in Figure 2(a) is on a log scale.) For values of ϕ such as .50 or less, the variance reductions are small, no more than about 7% when $v_i/\sigma_u^2 = 1$, and less as v_i/σ_u^2 moves away from 1.0. Large variance reductions require larger values of ϕ. For example, to achieve a 20% or greater reduction in variance requires $\phi \geq .75$.

Esteban et al. (2012) reported results related to those of Figure 2(a) obtained from a simulation study of the FH-AR(1) model, though augmented with a time-invariant area level random effect. This feature, and some other differences (most notably that their simulations provide estimates of the full prediction MSEs, not just a first order approximation) make their specific numerical results not directly comparable to ours. However, their results obtained with the alternative values of $\phi = 0, .25, .5, .75$ (denoted as ρ in their paper) are consistent with the general conclusions we draw from Figure 2(a). First, they found that borrowing from past data yielded little if any benefit for $\phi \leq .5$. Then, for $\phi = .75$, their augmented FH-AR(1) model appears (judging from their Figure 4.1) to reduce prediction MSEs by about

10%, or in some cases slightly more, relative to those obtained by applying this model with ϕ fixed at 0. Their simulation model assumed different values of the sampling variances across areas and time points, resulting in values of their ratio of sampling to model variance ranging from roughly .325 to .8 overall. The values within this range that were covered by a given simulation experiment varied as this depended on the value used for ϕ in the experiment. Esteban et al. also remarked that other simulations they did without the time-invariant random effect led to the same basic conclusions.

Figure 2(b) shows contours of percentage variance reductions from using the AR(1) model versus the bivariate model when the latter is applied to current year survey estimates and the average of survey estimates over the previous five years. We take this use of the five-year average as an approximation to the use of ACS 5-year estimates. Since the calculations assume the AR(1) is the true model, the bivariate model must have higher posterior variances. However, the reduction in variance from using the AR(1) model is generally small – less than 10% except for a small region in the upper left corner of the plot for high values of ϕ and values of $v_i/\sigma_u^2 < .5$.

One might wonder whether larger variance improvements from the AR(1) or bivariate models might result if more years of data were used compared to the six years assumed for the plots of Figure 2. Doing the same contour plots for the cases of 10 years of data and 20 years of data produced little change in the plots, except for very large values of ϕ and within a limited range of large v_i/σ_u^2 values, where more substantial advantages to the AR(1) over the bivariate model were observed. Over almost all of the range of ϕ and v_i/σ_u^2, using more years of past data appeared to make little difference.

The values of the variance ratios, v_i/σ_u^2, across the areas $i = 1,\ldots,m$ in the model will clearly affect how much variance improvement is achieved in specific areas. To gauge this effect for our application, we fitted a linear FH model to the ACS estimated county poverty rates, for which we had the sampling variances v_i from the GVF, and, using the posterior mean of σ_u^2, we calculated the ratios v_i/σ_u^2. Figure 3 shows a histogram of these variance ratios with the x-axis on a log scale. Most of the values lie between .1 and 10, though some extend beyond this. The variance ratios across the U.S. counties thus reflect much of the x-axis range of the contour plots in Figure 2.

Figures 2 and 3, the simulation results of Esteban et al. (2012), and the estimates of ϕ for the AR(1)-BLN model, suggest that for our application only small improvements in posterior prediction variances would be realized from the AR(1) or bivariate models compared to the univariate BLN. This is consistent with the

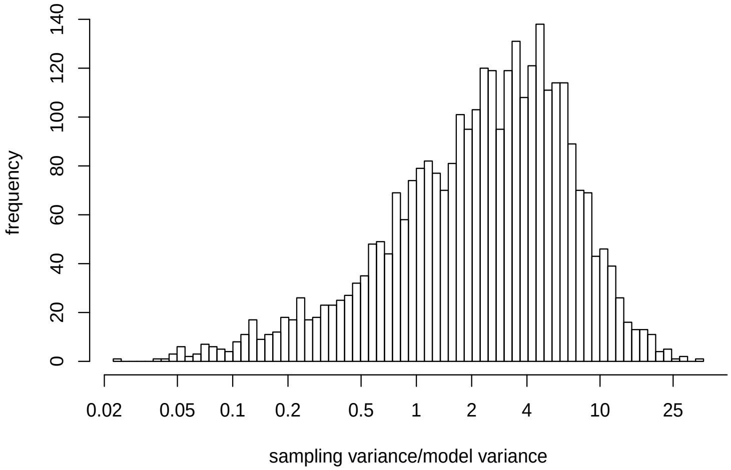

Figure 3: Histogram of the ratios of the sampling variances to the model variance in the FH model for the 2012 U.S. county poverty rates of school-aged children

posterior variance comparisons discussed in Section 3.1. Two other results from these comparisons may still seem surprising. First, the improvements for the bivariate BLN model are somewhat larger than the theoretical calculations for the linear model would suggest. Second, the improvements for the bivariate BLN model are larger than are those for the AR(1) model. While one would expect some limitations on how well calculations for linear FH models with parameters assumed known apply to fitted BLN models, that does not seem to explain these results since we obtained very similar results when we made the same comparisons using the bivariate and AR(1) extensions to the FH model applied to county poverty rates. In this case the bivariate FH model reduced prediction error standard deviations and variances compared to the univariate FH model by, on average, 5% and 9% (compared to 5% and 11% for the bivariate BLN). Corresponding figures for the AR(1) FH model were 2.7% and 5.2% (compared to 2.3% and 4.6% for the AR(1) BLN). In any case, differences between the comparisons for the BLN models and those for the FH model (both empirical and theoretical results) are not large, and all lead to the main conclusion that, given the value of ϕ for the AR(1) model, modest variance reductions would be achieved by the bivariate or AR(1) models relative to the univariate model.

3.3. Impact of removing model covariates

To illustrate a case where greater improvement would be expected from borrowing information over time, we repeated our empirical analyses after removing the regression covariates from the BLN models, leaving only the intercept terms. Without the regressors, the posterior means of ρ and ϕ skyrocketed to .92 and .94, respectively. We are now in the region of the parameter space where, by Figure 2(a), we would expect to see very substantial reductions in posterior variances from using a bivariate or AR(1) model rather than a univariate model. For this case, Figure 4 shows substantial differences between both the posterior means and posterior standard deviations of county poverty rates from the univariate and bivariate BLN models. In fact, we now see an average 25% reduction in posterior standard deviations and a 43% reduction in posterior variances from using the bivariate versus the univariate model. The AR(1) and bivariate BLN models performed similarly (results not shown on the plots), with the AR(1) yielding, on average, 1.3% higher posterior standard deviations compared to the bivariate BLN. The average reductions in standard deviations and variances for both the bivariate and AR(1) FH models for poverty rates were 26% and 45%.

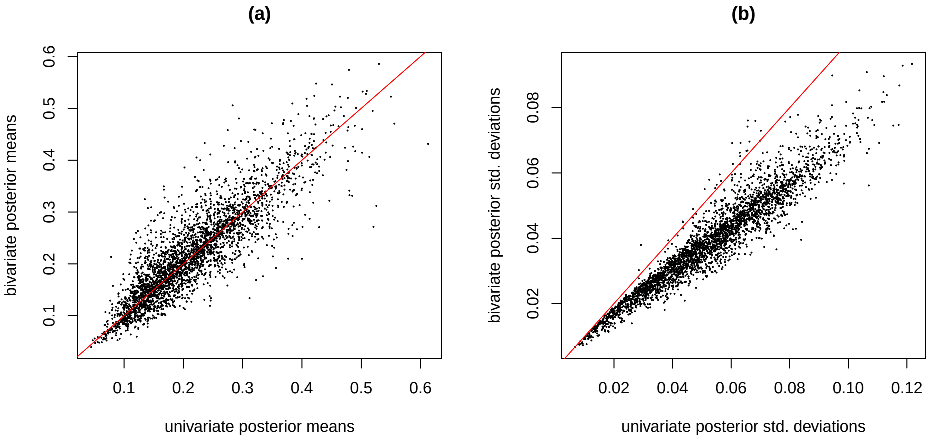

Figure 4: Comparisons of the posterior means and standard deviations for the 2012 U.S. county poverty rates of school-aged children for the univariate and bivariate BLN models with no regressors

3.4. Some model checks

For the linear (FH) model, where $y_i = (x_i'\beta + u_i) + e_i$, examination of standardized residuals defined as $(y_i - x_i'\hat{\beta})/[\widehat{\text{var}}(y_i - x_i'\hat{\beta})]^{1/2}$ provides a standard model check. We seek an analog for the BLN model (1)–(2). Since the inverse to (2) is $p_i = (1 + e^{-(x_i'\beta + u_i)})^{-1}$ and $E(u_i) = 0$, it may seem natural to use residuals defined as

$y_i/n_i - \hat{p}_i$ where $\hat{p}_i = (1+e^{-x_i'\hat{\beta}})^{-1}$ and $\hat{\beta}$ is an estimate of β. However, even with β known, $(1+e^{-x_i'\beta})^{-1}$ is not an unbiased estimator of $E(y_i/n_i)$ due to the nonlinearity of the logistic transformation and the presence of the random effects u_i. Instead, we define residuals as $y_i/n_i - E(y_i)/n_i$ and compute

$$E(y_i/n_i) = (1/n_i)E_{p_i}[E(y_i|p_i)] = E_{p_i}(p_i) = \int_{-\infty}^{\infty}(1+e^{-z_i})^{-1}f(z_i)dz_i \tag{10}$$

where $z_i = \text{logit}(p_i)$, $f(z_i)$ is the $N(x_i'\beta, \sigma_u^2)$ density, and $E_{p_i}(\bullet)$ denotes unconditional expectation over the distribution of p_i.

To standardize the residuals we need the unconditional variance

$$\begin{aligned}\text{var}(y_i) &= E_{p_i}[\text{var}(y_i|p_i)] + \text{var}_{p_i}[E(y_i|p_i)]\\ &= E_{p_i}[n_ip_i(1-p_i)] + \text{var}_{p_i}[n_ip_i]\\ &= n_iE_{p_i}[p_i] - n_iE_{p_i}[p_i^2] + n_i^2\text{var}_{p_i}[p_i]. \end{aligned}\tag{11}$$

To compute (11) requires computing $E_{p_i}[p_i^2]$ which, analogous to (10), is

$$E_{p_i}[p_i^2] = \int_{-\infty}^{\infty}(1+e^{-z_i})^{-2}f(z_i)dz_i. \tag{12}$$

Substituting the posterior means of β and σ_u^2 into $f(z_i)$, both (10) and (12) can readily be computed by numerical integration. We used the "integrate" function in R (R Core Team 2013) for this purpose. We then computed standardized residuals as $[y_i/n_i - E(p_i)]/[\text{var}(y_i)^{1/2}/n_i]$.

Figure 5 plots such standardized residuals for 2012 from the equation for $\tilde{y}_{1i}$ of the bivariate BLN given by (4)–(5) against county effective sample sizes $\tilde{n}_{1i}$. (We could equally well do this for residuals from the equation for $\tilde{y}_{2i}$, but focus here on checking the model for $\tilde{y}_{1i}$ since our interest lies in predictions of p_{1i}.) For $\tilde{n}_{1i}$ "sufficiently large", standard normal distribution inferences (e.g., ±2.57 for a 99% confidence interval, as denoted by the blue dashed lines on the plot) may be appropriate given the approximate normal distribution of the binomial, although precisely how large $\tilde{n}_{1i}$ must be for this approximation to hold is unclear (Brown, Cai, and DasGupta 2001). In any case, in the plot the bulk of the residuals look reasonably symmetrical, with no systematic biases related to sample size (which is strongly related to population size). There are a number of large positive residuals, though mostly these occur at the smaller effective sample sizes, especially for $\tilde{n}_i$ of about 30 or less, where the direct estimates are erratic. It may seem odd that there is not a corresponding set of large in magnitude negative residuals. This is due to the fact that $\tilde{y}_i/\tilde{n}_i \in [0,1]$ while all the predicted p_i values are less than 0.54. Extreme

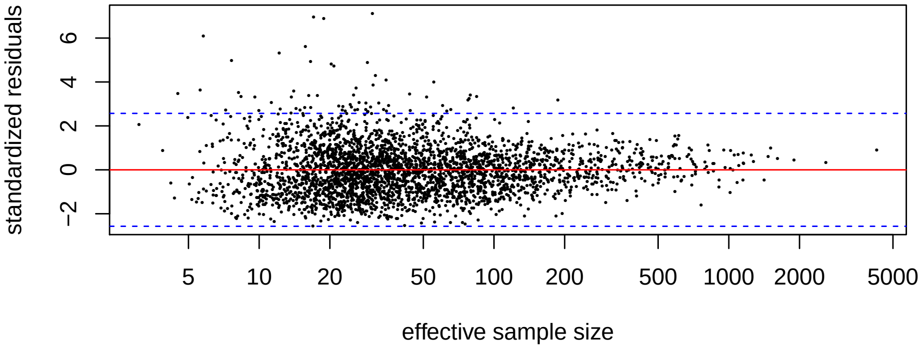

Figure 5: Standardized residuals from the 2012 bivariate BLN model's equation for y_{1i} plotted against county ACS effective sample sizes.

negative residuals are thus unlikely, while extreme positive residuals occur when $\tilde{y}_i/\tilde{n}_i$ is large, even 1.0, as happens sometimes with small samples.

We also examined a plot (not shown) of the standardized residuals against the predicted p_i values, which mimics a standard regression diagnostic (plot residuals against fitted values). This plot did not suggest any systematic biases related to the predicted county poverty rates.

Brown et al. (2001) suggest as a "calibration diagnostic" comparing model predictions aggregated to larger areas against corresponding direct survey estimates. In SAIPE production the county model predictions of the number of age 5–17 children in poverty are raked (rescaled) to force agreement with corresponding state estimates obtained from an FH model applied to direct ACS estimates of state poverty rates. For large states substantial weight is given to the direct ACS estimate in the model predictions, and this raking is then similar to raking to the direct estimates. In any case, there is practical interest in how much raking of the county model predictions is required. We examine this here for the bivariate BLN and (unraked) SAIPE production county model predictions derived from the 2012 ACS data.

To explain this in more detail, for the bivariate BLN model we expand our notation slightly to let $\hat{p}_{ji}$ be the bivariate BLN county model prediction of the age 5–17 poverty rate for county i in state j (treating the District of Columbia (DC) as both a county and a state in this analysis), and N_{ji} be the 5–17 population estimate for county i obtained from the Census Bureau's population estimates program. (Actually, slight modifications are made of the N_{ji} to estimate the county "poverty universes", which exclude a relatively small set of persons for whom poverty status cannot be determined (Bell et al. 2015).) The predicted number of age 5–17 children in poverty for state j implied by its county model predictions is then $\sum_{i\in j}\hat{p}_{ji}N_{ji}$. The SAIPE production county model is an FH model for logarithms of the number of children age 5-17 in poverty (Bell et al. 2015). Predictions from this model are transformed to the original (unlogged) scale using a bias adjustment based on properties of the lognormal distribution and accounting for uncertainty due to estimating regression parameters of the model. These predictions are then simply summed across counties to yield state level predictions of the number in poverty.

Figure 6 plots percent differences of the state total estimates of the number of age 5-17 children in poverty from the two county models – bivariate BLN and SAIPE production – compared to the corresponding estimates derived from the SAIPE state model. The percent differences are defined as $100 \times (1 -$ SAIPE state model estimate/aggregated county model predictions) so positive values indicate aggregated county model predictions exceeding the state model predictions and negative values indicate aggregated county model predictions lower than the state model predictions. The percent differences are plotted for 50 states, with states sorted by their ACS sample sizes (number of addresses). We dropped Alaska because it contained 5 of the 6 counties omitted from the modeling due to their not being consistently defined for all years of our data, which prevented us from getting an implied state poverty prediction for Alaska from the bivariate BLN model. The other omitted county was in Texas, but it had inconsequential effects on the state total.

Somewhat greater percent differences are to be expected at the left of Figure 6 for the small states where the estimation uncertainty is highest. This tendency is apparent in the plot. Apart from this, if we examine the blue solid dots in the plot, we see that the percent differences for the bivariate BLN model appear to be usually no more than a few percent. The corresponding percent differences for the SAIPE production estimates (red circles) appear to usually exceed those from the bivariate BLN, as well as being generally larger in magnitude. These impressions are reflected by Table 1, which summarizes the distributions of the percent differences.

county model	min	1^{st} quartile	median	mean	3^{rd} quartile	max
Bivariate BLN	-6.3	-1.2	$-.5$	$-.3$	.8	3.4
SAIPE production	-4.3	.2	2.2	1.8	4.5	11.6

Table 1: Distribution (omitting Alaska – see text) of percent differences between the state aggregates of county model predictions of 5–17 in poverty from the bivariate BLN and SAIPE production models compared to the SAIPE state model estimates for 2012.

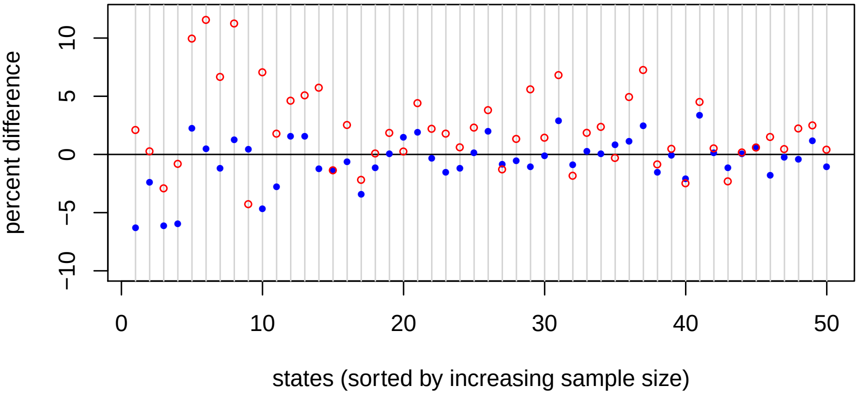

Figure 6: Percent differences between aggregated county model predictions of 2012 state total numbers of age 5–17 children in poverty and corresponding SAIPE state estimates. Red circles = SAIPE production model (unraked predictions); blue solid dots = bivariate BLN model.

Ideally we should take account of statistical uncertainty in the state level percent differences, but this is complicated, particularly for the bivariate BLN, by the dependence between the state and county model predictions due to both coming from models fitted to ACS data. As a conservative indication, 90% prediction intervals for the SAIPE state model predictions, expressed in multiplicative percentage terms, range from lows of around ±1.7% for the largest states (California and Texas) to highs of about ±11% to ±13% for some of the smallest (Wyoming, New Hampshire, and DC). These figures should overstate the uncertainty in the percent differences since we would expect positive dependence between the state and county model predictions.

4. Conclusions

Several conclusions stand out from the empirical and theoretical results presented in this paper. A general conclusion is that to achieve substantial variance reductions by jointly modeling current and past data requires fairly high levels of dependence over time in the random effects (model errors) of small area models. With modest levels of dependence, variance reductions from including past data are likely to be limited. A conclusion specific to the empirical example on modeling ACS poverty estimates is that the regression covariates used in the models do a good job explaining variation in poverty across counties and over time, leaving residuals with modest levels of dependence. Without these covariates in the models, the dependence over time in the model errors is strong, and borrowing information from past data then substantially reduces posterior (prediction error) variances.

A second general conclusion is that a bivariate model for the current year's estimate and the average of the estimates for some number of immediately preceding years may do about as well as an AR(1) model in borrowing information from past data for small area predictions. In fact, in the example bivariate models did slightly better than the corresponding AR(1) models. Additional comparisons could be made to models with more general dependence structures, such as a higher order AR model or a general 6×6 covariance matrix. While we intend to pursue this, we are confident that this will not alter the main conclusions expressed in the preceding paragraph. We also conjecture that bivariate models may do reasonably well in comparisons to other time series models with stationary autocorrelations, such as higher order AR models. It seems less clear whether this will be the case for models with nonstationary dependence, such as random walks. Consideration of the bivariate model is natural for the SAIPE application given that the ACS annually produces 5-year estimates for all U.S. counties and other small areas, and these 5-year pooled sample estimates can be thought of as similar to 5-year averages of 1-year estimates. While the bivariate model may seem less natural in other applications, it could be considered as a somewhat simpler alternative to using a time series model.

Appendix: Calculating Prediction MSEs for the Bivariate and FH-AR(1) Models

For extending the linear FH model to bivariate and AR(1) versions, let y_{it} be the direct survey estimate for area i and time $t = 1, \ldots, T$ of population characteristic Y_{it}, so $y_{it} = Y_{it} + e_{it}$ where e_{it} is the sampling error. For simplicity we assume the model parameters are known (first order approximation) and also assume normality, so that the best linear predictor (BLP) is the conditional expectation and the prediction MSE is the conditional variance. With parameters assumed known we need not explicitly consider the regression mean for $E(Y_{it})$, as this does not affect the conditional variances, which are our focus here. Also, since the FH model assumes independence over areas i, the BLP for area i then uses data for only that area, so we simplify the notation by dropping the subscript i. We further simplify by assuming that $\text{var}(Y_t) = \text{var}(u_t) = \sigma_u^2$ and $\text{var}(e_t) = v$ are constant over time. Within this simplified setup, we seek MSEs for the bivariate and FH-AR(1) predictors of Y_T, the most recent true population quantity, given data $\mathbf{y} = [y_1, \ldots, y_T]'$.

Let $\mathbf{z} = A\mathbf{y} = A\mathbf{Y} + A\mathbf{e}$ where A is a $k \times T$ matrix of the form

$$A = \begin{bmatrix} A_{11} & \mathbf{0} \\ \mathbf{0}' & 1 \end{bmatrix} \quad \Rightarrow \quad \mathbf{z} = \begin{bmatrix} A_{11}\mathbf{y}_1 \\ y_T \end{bmatrix} \tag{13}$$

where $\mathbf{y}_1 = [y_1, \ldots, y_{T-1}]'$ and $\mathbf{0}$ is a $(k-1) \times 1$ vector of zeroes. For the FH-AR(1) model $k = T$ and $A_{11} = I_{T-1}$, while for the bivariate model $k = 2$ and $A_{11} = (T-1)^{-1}[1, \ldots, 1]$. Letting $\Sigma_e \equiv \text{var}(\mathbf{e})$, and similarly defining Σ_u, Σ_y, and Σ_z, we have $\Sigma_y = \Sigma_u + \Sigma_e$, $\Sigma_z = A\Sigma_y A'$, and $\text{cov}(\mathbf{e}, \mathbf{z}) = \Sigma_e A'$. From standard results on conditional variance in a multivariate normal distribution, and since predicting $\mathbf{e}$ is equivalent to predicting $\mathbf{Y}$, then

$$\text{var}(\mathbf{Y}|\mathbf{z}) \equiv \text{var}(\mathbf{e}|\mathbf{z}) = \Sigma_e - \Sigma_e A'(A\Sigma_y A')^{-1} A\Sigma_e.$$

We are assuming $\Sigma_e = vI$, and we write $\Sigma_u = \sigma_u^2 R$, where R is the $T \times T$ correlation matrix of $\mathbf{Y}$. Then $\Sigma_y = \sigma_u^2(\lambda I + R)$ where $\lambda = v/\sigma_u^2$ is the noise-to-signal ratio. Thus,

$$\begin{aligned} \text{var}(\mathbf{e}|\mathbf{z}) &= vI - vA'[\sigma_u^2 A(\lambda I + R)A']^{-1}Av \\ &= v\{I - \lambda A'[A(\lambda I + R)A']^{-1}A\}. \end{aligned}$$

Let $\Omega \equiv [\omega_{j\ell}] = [A(\lambda I + R)A']^{-1}$. We are interested in the (T,T)th element of $\text{var}(\mathbf{e}|\mathbf{z})$, which is

$$\text{var}(Y_T|\mathbf{z}) \equiv \text{var}(e_T|\mathbf{z}) = v\left\{1 - \lambda[\mathbf{0}', 1]A'\Omega A\begin{bmatrix}\mathbf{0} \\ 1\end{bmatrix}\right\}.$$

From the definition of A in equation (13), $[\mathbf{0}', 1]A' = [\mathbf{0}', 1]$, so that this reduces to

$$\text{var}(Y_T|\mathbf{z}) = v(1 - \lambda\omega_{TT}) \tag{14}$$

where ω_{TT} is the (T,T)th element of Ω. The expression (14) is easily computed given v, σ_u^2, and R. For our comparisons, R is the AR(1) correlation matrix given in equation (9), which is determined solely by ϕ. Hence, Ω is determined by λ and ϕ. Note that for the bivariate model, A is $2 \times T$ and Ω is then a 2×2 matrix.

The prediction MSE of the univariate FH model is $\text{var}(Y_T|y_T) = \sigma_u^2 v/(\sigma_u^2 + v)$ (Rao and Molina 2015, eq. (6.1.8)). The percent reduction in prediction MSE from the FH-AR(1) or bivariate models relative to the univariate FH model is thus 100 times

$$\begin{aligned} 1 - \frac{\mathrm{var}(Y_T|\mathbf{z})}{\mathrm{var}(Y_T|y_T)} &= 1 - \frac{\sigma_u^2 + v}{\sigma_u^2 v} v(1 - \lambda \omega_{TT}) \\ &= 1 - (1 + \lambda)(1 - \lambda \omega_{TT}). \end{aligned}$$

This expression depends on only λ and ϕ.

Acknowledgements

We thank Roderick Little, Hal Stern, and Alan Zaslavsky who, in a 2007 review of SAIPE models, suggested investigation of binomial models using effective sample sizes for modeling ACS poverty estimates.

Disclaimer

This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

REFERENCES

BELL, W. R., BASEL, W. W., CRUSE, C., DALZELL, L., MAPLES, J. J., O'HARA, B., POWERS, D. (2007). Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties. SAIPE technical report, U.S. Census Bureau, URL http://www.census.gov/did/www/saipe/publications/files/report.pdf.

BELL, W. R., BASEL, W. W., MAPLES, J. J., (2015). An Overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) Program. In *Analysis of Poverty Data by Small Area Methods*, Monica Pratesi (ed.), London: Wiley, to appear.

BOX, G. E. P., JENKINS, G. M. (1970). *Time Series Analysis: Forecasting and Control*, San Francisco: Holden Day.

BROWN, L. D., CAI, T. T., DASGUPTA, A., (2001). Interval Estimation for a Binomial Proportion, *Statistical Science*, 16, 101-133.

BROWN, G., CHAMBERS, R., HEADY, P., HEASMAN, D. (2001). Evaluation of Small Area Estimation Methods – An Application to Unemployment Estimates from the UK LFS. In *Proceedings of the Statistics Canada Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*.

DATTA, G. S., LAHIRI, P., MAITI, T., LU, K. L. (1999). Hierarchical Bayes Estimation of Unemployment Rates for the States of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

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.

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, 269–77.

FRANCO, C., BELL, W. R. (2013). Applying Bivariate Binomial/Logit Normal Models to Small Area Estimation. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 690–702, URL http://www.amstat.org/sections/srms/Proceedings/.

GHOSH, M., NATARAJAN, K., STROUD, T.W.F., CARLIN, B.P. (1998). Generalized Linear Models for Small Area Estimation. *Journal of the American Statistical Association*, 93, 273–282.

GHOSH, M., NANGIA, N., KIM, D. H. (1996). Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach. *Journal of the American Statistical Association*, 91, 1423-1431.

HAWALA, S., LAHIRI, P. (2012). Hierarchical Bayes Estimation of Poverty Rates. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 3410–3424, URL http://www.census.gov/did/www/saipe/publications /files/hawalalahirishpl2012.pdf.

HUANG, E. T., BELL, W. R. (2012). An Empirical Study on Using Previous American Community Survey Data Versus Census 2000 Data in SAIPE Models for Poverty Estimates. Research Report Number RRS2012–4, Center for Statistical Research and Methodology, U.S. Census Bureau, URL http://www.census.gov/srd/papers/pdf/rrs2012–04.pdf.

LIU, B. M., LAHIRI, P., KALTON, G. (2007). Hierarchical Bayes Modeling of Survey Weighted Small Area Proportions. *Proceedings of the American Statistical Association*, Survey Research Section, 3181–3186.

PLUMMER, M. (2010). JAGS - Just Another Gibbs Sampler (JAGS 2.1.0., May 12, 2010). URL https://sourceforge.net/projects/mcmc-jags.

PRATESI, M., GIUSTI, C., MARCHETTI, S., SALVATI, N., TZAVIDIS, N., MOLINA, I., DURBÁN, M., GRANÉ, A., MARÍN, J. M., VEIGA, M. H., MORALES, D., ESTEBAN, M.D., SAŃCHEZ, A., SANTAMARÍA, L., MARHUENDA, Y., PÉREZ, A., PAGLIARELLA, M. C., RAO, J. N. K., FERRETTI, C. (2010). Small Area Estimation of Poverty and Inequality Indicators: Pilot Applications. Work Package 2 (D17) of Small Area Methods for Poverty and Living Conditions Estimates Project. URL http://www.sample-project.eu/SAMPLEwp2d17.pdf.

R CORE TEAM (2013). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/.

RAO, J. N. K., MOLINA, I. (2015). *Small Area Estimation* (2nd ed.), Hoboken, New Jersey: Wiley.

SAEI, A., CHAMBERS, R. (2003). Small Area Estimation Under Linear and Generalized Linear Mixed Models with Time and Area Effects. Southampton, UK, Southampton Statistical Sciences Research Institute. (S3RI Methodology Working Papers, (M03/15)). URL http://eprints.soton.ac.uk/8165/.

SLUD, E. V. (2000). Models for Simulation and Comparison of SAIPE Analyses. SAIPE Technical Report, U.S. Census Bureau, URL http://www.census.gov/did/www/saipe/publications/files/saipemod.pdf.

SLUD, E. V. (2004). Small Area Estimation Errors in SAIPE Using GLM versus FH Models. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 4402–4409, URL http://www.amstat.org/sections/srms/Proceedings/.

TACIAK, J., BASEL, W. W. (2012). Time Series Cross Sectional Approach for Small Area Poverty Models. *Proceedings of the American Statistical Association, Section on Government Statistics*, URL http://www.census.gov/did/www/saipe/publications/files/JTaciakBaseljsm2012.pdf.

U.S. CENSUS BUREAU (2014). *American Community Survey Design and Methodology (version 2.0, January 2014)*, URL http://www.census.gov/acs/www/methodology/methodology_main/.

YOU, Y. (2008). An Integrated Modeling Approach to Unemployment Rate Estimation for Sub-Provincial Areas of Canada. *Survey Methodology*, 34, 19-27.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 585–602

SMALL AREA ESTIMATES OF THE POPULATION DISTRIBUTION BY ETHNIC GROUP IN ENGLAND: A PROPOSAL USING STRUCTURE PRESERVING ESTIMATORS

Angela Luna, Li-Chun Zhang[1], Alison Whitworth, Kirsten Piller[2]

ABSTRACT

This paper addresses the problem of producing small area estimates of Ethnicity by Local Authority in England. A Structure Preserving approach is proposed, making use of the Generalized Structure Preserving Estimator. In order to identify the best way to use the available aggregate information, three fixed effects models with increasing levels of complexity were tested. Finite Population Mean Square Errors were estimated using a bootstrap approach. However, more complex models did not perform substantially better than simpler ones. A mixed-effects approach does not seem suitable for this particular application because of the very small sample sizes observed in many areas. Further research on a more flexible fixed-effects estimator is proposed.

1. Introduction

Estimates of demographic characteristics are among the main outputs of National Statistical Institutes (NSIs). In addition to national and regional estimates, for topics such as Labour Force, Household composition or Ethnicity, periodic estimates at lower levels of geographic aggregation are in high demand both for public policy and research purposes.

In census years, given the availability of data for almost all individuals in the population, it is straightforward to produce reliable estimates for small geographic domains. In contrast, during the inter-censal period, updated socio-demographic data can only be obtained via sample surveys or administrative systems. It is generally difficult to obtain reliable direct estimates for small geographic domains from sample surveys due to the small sample sizes. Data from administrative systems do not have this problem but, in contrast, may not cover the topics of interest. Moreover, definitions of the variables and domains in administrative

[1] University of Southampton.
[2] Office for National Statistics ONS-UK.

sources reflect the requirements of the administrative systems, which may be different from those for statistical purposes. This can result in comparability issues with figures obtained from population censuses or household surveys.

From the statistical perspective, the estimation for small domains in the presence of limited, or even null domain-specific sampling data can be framed within the field of Small Area Estimation (SAE). Research in SAE has gained relevance in the last decades due to an increasing demand for small area outputs in the Official Statistics sector, as well as in many others. Readers interested in SAE can find a comprehensive account of methods in Rao (2003). For a review of the most important developments of the last decade, see Pfeffermann (2013).

Implementation of SAE methods in the field of Official Statistics faces specific challenges and specialized research has been encouraged in the European context. Projects such as EURAREA (Eurostat, 2001-2004) and EU-SAE (ESSnet, 2009-2012) provide comprehensive reviews of the available SAE methods with potential applications in a broad set of topics covered by Official Statistics, taking into consideration the specific requirements and characteristics of European statistical systems. Special attention has been given to the use of SAE methods for the measurement of poverty, via collaborative projects such as SAMPLE (Small Area Methods for Poverty and Living conditions Estimates, European Commission, 2008-2011) or AMELI (Advanced Methodology for Laeken Indicators, 2008-2011). The deliverables of all the above mentioned projects are available online.

At present, comparatively few official figures in the region are being produced using SAE methods. In the UK case, the Office for National Statistics (ONS) periodically disseminates small area estimates regarding three main topics: population estimates by age and sex using the Census and its coverage survey; average household income and households in poverty using the Family Resources Survey and administrative data maintained by the Department for Work and Pensions; and unemployment, making use of the Annual Population Survey and the administrative register for Jobseeker’s Allowance.

Nonetheless, the interest in understanding the potential gains that can be obtained from a more extensive use of SAE methods in the context of official statistics remains. The ONS established the Census Transformation Programme in January 2015 to take forward the National Statistician’s recommendation to make the best use of all available data in the production of population statistics. This involves research into the potential use of administrative data as well as surveys to produce population, household and characteristic information currently provided in the Census. SAE methods provide a framework for intergrating sources. In this paper we investigate the problem of how to obtain estimates of the distribution of the population by ethnic group, in each Local Authority (LA) of England, using proxy and survey data. Such estimates are required by local and central government for planning services and formulating policy. More generally researchers, local authorities, health authorities and other public and private sector organisations could use them to gain an up-to-date picture of the ethnic composition of local populations and to monitor diversity and anti-discrimination programmes.

Ethnicity is a variable for which the use of Structure Preserving Estimators (SPREE) (Purcell and Kish, 1980) seems natural. Most SAE methods combine existing survey data for the variable of interest with relevant covariate information obtained from censuses or administrative sources, to obtain better estimates than those from the survey alone. For Labour Force status for instance, covariates such as sex, age or level of education can provide some explanatory power, see Molina et al. (2007) and Scealy (2010). In the case of ethnicity, on the other hand, it is difficult to identify such a set of covariates. Instead, for post-censal updates of the LA by ethnicity distribution, the corresponding aggregated census table can always be treated as a proxy for the table of interest.

When a proxy is available, the SPREE approach allows for an intuitive modelling of the relationship between the so-called association structure, or simply the structure of both the proxy table and the table of interest. The SPREE approach is particularly compelling in the case where the margins of the table of interest are known in advance or can be accurately estimated because, given the margins, the structure is the only unknown component to be estimated. This will be explained in more detail in Section 2.

This application addresses the particular problem of obtaining updated census tables of LA by ethnicity during the inter-censal period. However, it is important to notice that population censuses in general are going through a process of redesign in many European countries. More emphasis is being given to alternative operations based on demographic systems that use information from administrative sources alone or in combination with survey data. In such a context, the potential impact of SAE methods, including the SPREE approach and its extensions, is expected to increase considerably in the future.

The rest of the paper is organised as follows. In the next section, the underpinning idea behind the SPREE approach and the GSPREE extension (Zhang and Chambers, 2004) is discussed in more detail. Section 3 describes the characteristics of the empirical exercise performed to obtain estimates of the distribution by ethnic group and LA in England. Section 4 presents the results of our analysis. Finally, Section 5 discusses the main results and points out some topics for future work.

2. SPREE approach

2.1. Structure Preserving Estimator (SPREE)

Denote by Y the population table of interest, with cells Y_{aj}, where $a = 1,\ldots,A$ indexes the set of areas and $j = 1,\ldots,J$ indexes the categories of the variable. Define $\zeta_{aj}^{Y} = \log Y_{aj}$. Y can be represented in the form of a saturated log-linear model as:

$$\zeta_{aj}^{Y} = \alpha_{0}^{Y} + \alpha_{a}^{Y} + \alpha_{j}^{Y} + \alpha_{aj}^{Y}, \qquad (1)$$

where $\alpha_{0}^{Y} = \overline{\zeta_{\bullet\bullet}^{Y}}$ (the dot indicating summing over the respective subscript), $\alpha_{a}^{Y} = \overline{\zeta_{a\bullet}^{Y}} - \alpha_{0}^{Y}$, $\alpha_{j}^{Y} = \overline{\zeta_{\bullet j}^{Y}} - \alpha_{0}^{Y}$ and $\alpha_{aj}^{Y} = \zeta_{aj}^{Y} - \alpha_{0}^{Y} - \alpha_{a}^{Y} - \alpha_{j}^{Y}$, for $a = 1,\ldots,A$, $j = 1,\ldots,J$. Following Purcell and Kish (1980), equation (1) can be used to decompose Y into two parts: the *association structure* and the *allocation structure*. The former corresponds to the terms $\left\{\alpha_{aj}^{Y}\right\}$, also called *interactions*, and determines the relationship between rows and columns in the table. In the theoretical case where rows and columns are independent, all the interaction terms are zero. The latter, given by the terms α_{0}^{Y}, $\left\{\alpha_{a}^{Y}\right\}$and $\left\{\alpha_{j}^{Y}\right\}$, carries information about the scale of the table and the disparities within the sets of rows and columns and is implicitly determined by the row and column margins of the table.

Notice that in the SAE setting, it is easier to obtain information related to the allocation structure than to the association structure. Even if Y remains unknown, accurate estimates of the row marginal, i.e. the area sizes, can be obtained either from administrative sources or from population estimates. Similarly, given that the column marginal corresponds to the aggregation over the entire set of areas, it can usually be accurately estimated using survey data, if not available from other sources.

Given the margins of Y, i.e., its allocation structure, a proxy of the table of interest, denoted by X, can be used to estimate the association structure of Y. The term proxy is used here in the customary sense of *proxy variable* as defined in Upton and Cook (2008): "A measured variable that is used in the place of a variable that cannot be measured". A proxy table is therefore supposed to contain information for the same set of areas and regarding a similar characteristic as the table of interest. In particular, it is assumed to have the same dimension $A \times J$. Notice that for demographic characteristics during inter-censal periods, the corresponding tables from the census year are obvious proxies. More generally, proxies can be derived not only from censuses but also from administrative sources.

For the two-way case, the SPREE of Purcell and Kish (1980) simply uses the association structure of the proxy table as an estimate for the association structure of the table of interest. In other words, denoting by $\left\{\alpha_{aj}^{X}\right\}$ the interaction terms for the proxy table X defined as in equation (1), the SPREE is characterised by the *structural equation*: $\alpha_{aj}^{Y} = \alpha_{aj}^{X}$, for $a = 1,\ldots,A$, $j = 1,\ldots,J$.

The procedure proposed by Purcell and Kish (1980) to obtain the SPREE of Y is straightforward. The known margins of Y are imposed on X using a multiplicative raking procedure such as the Iterative Proportional Fitting (IPF) algorithm (see for instance Agresti, 2013, p. 365-366). This ensures that the association structure of the estimated and proxy tables are the same. Fitting a saturated log-linear model with an offset term given by the interactions α_{aj}^{X} is an alternative way to obtain the same estimate (Noble et al., 2002).

However, assuming that the proxy and the table of interest share exactly the same association structure is clearly restrictive in practice. Other estimators have been proposed to *preserve* in a more flexible way the association structure, leading to what we have called the SPREE approach. The modifications to the initial SPREE of Purcell and Kish (1980) go in two main directions: i) by relaxing the structural equation of SPREE to consider other types of relationship between the two association structures and ii) by including cell-specific random effects. Besides the SPREE, the following estimators can be framed within this approach: the Generalized Structure Preserving Estimator (GSPREE, Zhang and Chambers, 2004), the Extended Structure Preserving Estimator (ESPREE, Cinco, 2010) and the estimator proposed in Berg and Fuller (2014). Notice that in all the above mentioned cases the allocation structure is imposed by benchmarking the estimates to a set of known margins. The benchmarking has the additional advantage of providing some degree of protection against misspecification of the assumed model (Pfeffermann, 2013).

2.2. Generalized Structure Preserving Estimator (GSPREE)

In some cases, it is possible to have access to a survey estimate of Y. Notice that the small area problem persists because the direct estimates of the cell totals are usually too unstable to be useful, due to small sample sizes. The GSPREE (Zhang and Chambers, 2004) proposes to use such information to *update* the association structure of the proxy table, aiming to reduce the bias of the SPREE. The GSPREE is characterised by the structural equation $\alpha_{aj}^{Y} = \beta\alpha_{aj}^{X}$ for $a=1,\dots,A,\ j=1,\dots,J$. Clearly, the SPREE corresponds to the particular case $\beta=1$.

An estimation procedure for β built directly from the structural equation involves several problems. Small sample sizes can lead to zero survey estimates for some of the cells, in which case the interaction terms for the survey estimate of Y are not defined. Moreover, even if all cells have a positive estimate, there is not a *natural* distribution that can be assumed for the interactions – as there is for the proportions or the counts – making it difficult to justify a standard approach such as Maximum Likelihood, for instance.

Therefore, instead of formulating a model in the interaction scale, Zhang and Chambers (2004) propose to estimate β using the Generalized Linear Structural Model (GLSM), a model relating the within-area proportions of the proxy table and the table of interest, on the log scale centred around the average of the area. The equation that defines the GLSM is:

$$\eta^Y_{aj} = \lambda_j + \beta\eta^X_{aj} \tag{2}$$

where $\eta^Z_{aj} = \log\theta^Z_{aj} - J^{-1}\sum_k \log\theta^Z_{ak}$, $\theta^Z_{ak} = Z_{ak}\Big/\sum_l Z_{al}$ for $Z = X,Y$, and $\sum_j \lambda_j = 0$.

The terms in the decomposition given in equation (1) satisfy $\sum_j \alpha^Z_j = 0$ and $\sum_j \alpha^Z_{aj} = \sum_a \alpha^Z_{aj} = 0$ for $Z = X,Y$. Moreover, $\alpha^\theta_j = \alpha^Y_j$ and $\left\{\alpha^\theta_{aj}\right\} = \left\{\alpha^Y_{aj}\right\}$. Using these arguments it is straightforward to show that $\eta^Z_{aj} = \alpha^Z_j + \alpha^Z_{aj}$ for $Z = X,Y$, and therefore, that equation (2) is equivalent to the structural equation of the GSPREE. The λ_j are nuisance parameters with no practical interest.

The GLSM is fitted via Iteratively Weighted Least Squares (IWLS) using direct estimates of the within-area proportions $\hat{\theta}^Y_{aj}$ and estimates of their variances. By doing so, it is implicitly assumed that the structural equation of the GSPREE holds for the table of direct estimates as well, or at least, that the value of β that better relates the table of interest and the proxy table does not change when the former is substituted by its direct estimate. Once the estimate $\hat{\beta}$ has been obtained, the GSPREE of Y is calculated by imposing the known row and column margins on the table of exponentiated estimated interactions $\tilde{Y}_{aj} = e^{\hat{\beta}\alpha^X_{aj}}$, using IPF.

In the absence of estimates of the variance of the direct estimators, it is also possible to obtain fully model-based estimates of β. One possibility, mentioned in Zhang and Chambers (2004), is to assume a multinomial distribution for the sampling cell counts in each area, and obtain an estimator of β using Maximum Likelihood (ML). Notice that this approach implicitly assumes that the sampling design of the survey is ignorable for Y. Otherwise, direct estimates of the proportions can be used instead of the observed proportions, assuming a multinomial distribution for the direct estimates of the cell totals.

Despite not being mentioned in Zhang and Chambers (2004), fully model-based estimates of β under the GSPREE structural assumption can also be obtained assuming a Poisson distribution for the sampling counts y_{aj}. It is straightforward to show that the equation:

$$\log Y_{aj} = \gamma_a + \lambda_j + \beta\alpha^X_{aj} \tag{3}$$

with $\sum_j \lambda_j = 0$ is also equivalent to the structural equation of the GSPREE. Both the γ_a and the λ_j terms for $a = 1,\dots,A$, $j = 1,\dots,J$, are nuisance parameters. It is possible to fit (3) in a standard software using log-linear models and obtain the corresponding ML estimator of β. As with the fitting using the GLSM, this approach assumes that the structural equation also holds for the table of sample counts.

In the application presented in Section 4 we followed a fully model-based approach in order to simplify the fitting process. By doing so, we can be incurring in a misspecification of the variance structure of the sampling errors. Nevertheless, using an argument similar to that for the generalised estimating equation approach in Liang and Zeger (1986), it is possible to show that in such a case the estimator of β, although not fully efficient, would remain unbiased.

3. Empirical exercise: distribution of the population by Ethnicity at LA level in England

An empirical exercise was conducted with the aim of producing small area estimates of the distribution of the population by ethnic group for each LA in England. Given that some of the sources of information used in this exercise are subject to disclosure control, it was necessary to perform all the data analysis in a Safe Room of the Virtual Microdata Laboratory (VML) of ONS. Thus, in accordance with ONS standards and the principles set out in the Code of Practice for Official Statistics, full account has been taken of requirements to safeguard confidentiality and uphold relevant data security standards. All the calculations hereby presented are the responsibility of the authors.

This section starts with a description of the data sources used: the proxy table, the table of survey estimates and the benchmark totals for the columns and row margins. A description of the variable of interest and the definition of categories across the different sources is then provided. Finally, the models that were involved in the fitting process are presented.

3.1. Sources of information

Proxy Information

Proxy information for the distribution of Ethnicity at the LA level can be obtained for England from several sources. For this empirical exercise, aggregate data from the 2011 Census and the English School Census[3] were used.

The 2011 Census provides estimates of the counts of persons and households who are defined as usual residents of England and Wales on the 27th March. The estimated coverage rate for persons in the 2011 Census was 93%. The observed counts were adjusted by over and undercount, taking into account the characteristics of individuals and households who were missed from the Census enumeration.

[3] Access to and use of information from the School Census is authorised by data sharing regulations i.e. Statistics and Registration Service Act 2007 (Disclosure of Pupil Information) (England) Regulations 2009.

The English School Census targets the population attending school in England and it is carried out every year. It mostly covers the population between 2 and 19 years old, with almost full coverage of children between the compulsory school ages of 5 and 15. The main school census in January collects information on the pupil's ethnicity, which is not asked about in the two other collection periods in June and August. Whereas state maintained schools and non-maintained special schools are included, independent schools are not covered. This can result in some differences between the population estimates for children in compulsory school age obtained from this and other sources.

As the English School Census only provides a good coverage for children between 5 and 15 years old, it could be said that for the empirical exercise there is one source of proxy information for individuals in the ages *0-4* and *16 or more*, and two sources for those *between 5 and 15 years*. In order to use the appropriate models for each age group, age-group specific Census tables of LA by Ethnic group were produced. Regarding the School Census, the empirical exercise hereby presented used information collected in January 2013.

Survey estimates

Most household surveys carried out by the ONS collect demographic data. For this empirical exercise, the Annual Population Survey (APS) is used for the updated estimates for the population by ethnic group. The APS contains detailed information on ethnicity, has the biggest sample size among the periodic surveys and, except for the Isle of Scilly, it includes information for all Local Authorities in England.

The APS is a household survey that is designed to provide information at a local level, on many demographic and socio-economic topics. The data sets are published quarterly (January to December; April to March; July to June; and October to September) and contain approximately 250,000 individuals. They contain the Labour Force Survey (LFS) data and the boost samples to the LFS. The boost for England is called the English Local LFS (ELLFS) and has been designed to give a minimum sample size of economically active individuals for each local education authority. The APS data set for England therefore consists of four successive quarters from the LFS, plus the ELLFS boost.

Both the LFS and the ELLFS use a rotational sampling design involving waves. For the LFS, a sample of households is interviewed quarterly for five waves, inducing an 80% of overlap between samples of consecutive quarters. For the ELLFS a sample is interviewed once a year for four waves. Notice that the households are included in the APS only the first time they are interviewed, so that each respondent only appears in the data set once. Non-private households (some communal establishments, armed forces accommodation, etc.) are excluded from the sampling frame. For England the households are sampled through the Royal Mail Postcode Address File (PAF) and the National Health Service (NHS) communal accommodation list. This empirical exercise uses the data corresponding to July 2012 – June 2013**.** The reference point is taken as the midpoint, so approximately the 31st of December 2012, which ties in with the School Census data.

As with the data from 2011 Census, a survey table from the APS survey was produced for each one of the age groups *0-4, 5-15* and *16 or more*.

Benchmark totals

Estimates of the LA population sizes can be obtained from the official mid-year population estimates. These estimates are produced using the cohort component method, which uses information on components of population change to update the most recent census population. The previous year's population estimate by sex, age and LA of usual residence is aged on by one year. Births within the 12 months to the reference date are added to the population and deaths are removed. The net flows of migration are accounted for internal (cross border and between LA) and international flows. There are also adjustments for special populations (armed forces and prisoners) who are not represented in the data sources used for the components of population change.

The 2012 and 2013 mid-year population estimates at LA level were used to calculate the row marginal. As the reference date of such estimates is 30th of June of the corresponding year, an average of the mid-year population estimates for 2012 and 2013 would provide an estimate of the population close to the 31st of December 2012, consistent with the reference period of the other sources involved in this exercise.

The direct estimates of the total population size by ethnic group, obtained from the APS at the national level, are used as the column benchmark totals in this exercise. Neither for the ethnic group nor the LA margins, a disaggregation by age group was considered.

3.2. Definition of the categories of the variable

The variable Ethnic group is collected in England in a very detailed way. The APS collects information regarding 18 subcategories of Ethnicity, grouped in 7 main categories: White, Mixed/multiple ethnic groups, Asian/Asian British, Black/African/Caribbean/Black British, Chinese, Arab and Other ethnic group. The 2011 Census uses 18 subcategories grouped in 5 main categories, with Chinese included within Asian and Arab within Other. Finally, the English School Census considers a classification similar to the one of the Census, except there is not a specific subcategory for Arab and Chinese is included as a subcategory within Other instead of within Asian.

To use a classification that is fully compatible with the three aforementioned sources, this empirical exercise uses the classification: White, Mixed/multiple ethnic groups, Asian/Asian British, Black/African/Caribbean/Black British, Chinese and Other.

3.3. Models

In order to produce an estimate for the table of interest using the GSPREE, only a proxy and a survey estimate of the table of interest, and the corresponding set of row and column margins are required. However, as it was described in the previous section, for the age group *5-15* two different sources of proxy information are available in this case. To study how to better use these sources, the following three models, with increasing level of complexity, were considered:

- **Model 1:** uses the 2011 Census as the only source of proxy information. Both the proxy and the survey tables are aggregated at the LA versus Ethnicity level, without considering the age group.
- **Model 2:** uses the 2011 Census as the only source of proxy information. The proxy and the survey tables are split by age group and an independent fitting is performed for each one of the three age groups mentioned above. The three estimates of the population counts are summed up to produce one estimate of the target table. The table of estimates after aggregating by age group is then benchmarked to the column and row margins.
- **Model 3:** uses both 2011 Census and the English School Census as sources of proxy information. In analogy to Model 2, an independent fitting is performed for each age group. Each one of those fittings goes through two steps:
    - **Step 1:** construction of an auxiliary structure that is a convex linear combination of the two available structures. The coefficient of the 2011 Census structure in the convex combination, denoted by δ, is found via numerical optimisation, as the value that minimizes the deviance of the fitting of the model defined by equation (3) for that particular age group.
    - **Step 2**: Estimation of the table of interest for that age group, using the survey data, the auxiliary structure built in step 1 and the GSPREE.

  As for Model 2, the three estimates of the population counts are summed up to produce one estimate of the population table. The benchmark of rows and columns is only applied over this last table estimate.

Notice that Model 2 is a particular case of Model 3 where $\delta = 1$. Therefore, a Likelihood Ratio Test (LRT) can be used as a diagnostic tool to compare their fitting. Given that Model 1 does not fit the three age groups independently, it is not possible to consider it nested in either of the other two models. However, an approximate Likelihood Ratio Test (LRT) between Model 1 and Model 2 is performed by approximating the former as a particular case of Model 2 with the same β in all age groups.

4. Results

Possibly due to the sampling design of the APS, no appreciable differences were observed in the within-LA distributions of ethnicity calculated from the sampling counts, or from direct estimates of the population counts. Therefore, sampling counts were used as input for the models. Poisson and Multinomial Likelihoods were used for the estimation of β, the latter being closer to a Simple Random Sampling design stratified by LA. The estimates of β obtained under the two distributions differ only at the third decimal point. Here we present only the results for the Poisson MLE.

The ethnicity variable has a very unequal distribution in the population. Aggregating the data of the 2011 Census for the areas under consideration, the category *White* is dominant with 85.42% of individuals, followed by *Asian* (7.10%), *Black* (3.48%), *Mixed* (2.25%), *Other* (1.03%) and finally *Chinese* (0.72%). How different LAs deviate from that global distribution can be observed in Figure 1. Notice that for categories *Asian* and *Black* it is possible to find some areas with proportions considerably higher than the global proportion. Moreover, notice that in such areas, non-white individuals are predominantly from one of the two above mentioned categories instead of evenly distributed. Meanwhile, for the categories *Mixed*, *Chinese* and *Other,* the proportions are uniformly low in all local authorities.

The actual sampling fractions of the APS in some LAs can be quite small. An implicit sampling fraction was calculated by dividing the observed sample size by the corresponding projected population total in each LA. This varies between 0.05% and 2.5%, with an average of 0.8%.

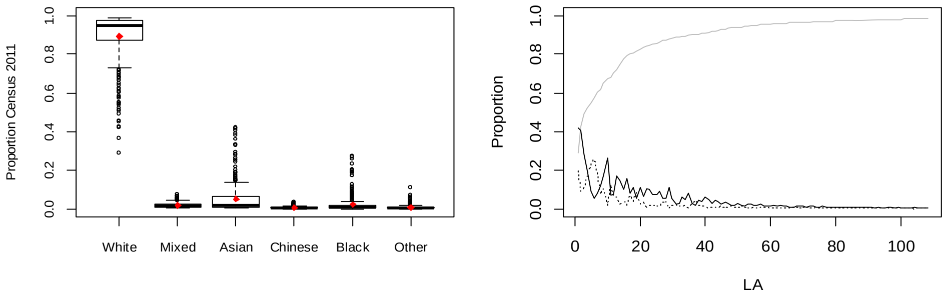

Figure 1. Distribution of Ethnicity by LA in the 2011 Census. (a) Boxplot proportions in each category by LA. Red diamond: mean. (b) Detail of the largest categories. Lines: *White*: continuous grey. *Asian*: dotted black. *Black*: continuous black. After sorting the LAs according to the proportion of *White,* one of each three LA was included in the plot

Given the low proportions of individuals belonging to categories such as *Chinese* or *Mixed*, as well as the small sample sizes observed in most LAs, some of the cells of the observed survey composition have zero observations, making it impossible to calculate their interaction terms directly. In principle, the presence of some sample zero cells does not necessarily cause a problem in terms of the estimation of the parameter of the GSPREE, when a fully model-based approach such as the one described in Section 2.2 is employed for this task. However, this means that the plausibility of the structural equation for this particular variable cannot be empirically checked using a scatterplot between interactions of the survey and proxy compositions. For illustration purposes, the pairs of interactions at the LA level for the 2001 Census and the 2011 Census in England are shown in Figure 2. Notice how, except for the category *Other*, interaction terms from the same composition 10 years before can still work fairly well as linear predictors. Unless period 2011-2013 behaves in a substantially different way than 2001-2011, it could be expected for the structural equation to hold at least approximately.

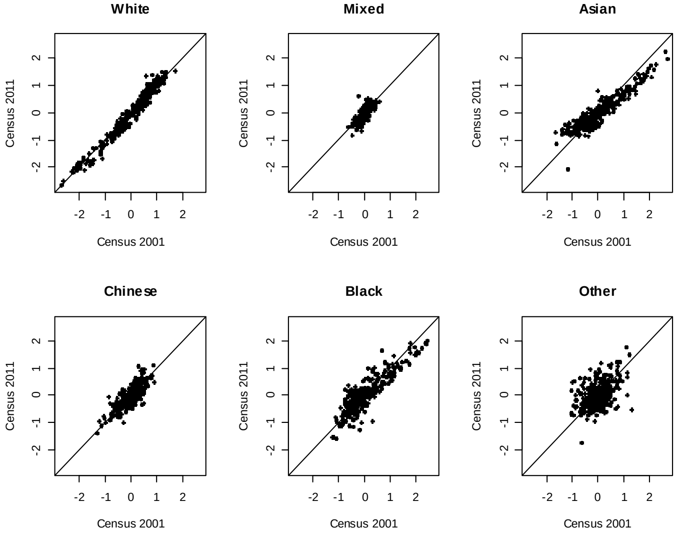

Figure 2. Interaction terms of the composition LA by Ethnicity , Census 2001 and 2011. Line: Y=X

The three fixed effects models stated in Section 3.4 were fitted to the data, using both the SPREE and GSPREE. In each case the estimated coefficient of the GSPREE estimator, β, is very close to 1, i.e. this estimator and the SPREE almost coincide. We therefore omit the results for the latter. The main results for the GSPREE are presented in Table 1. The last three columns contain the information to perform a LRT comparing the models in increasing order of complexity, as explained in section 3.4. In all cases there is evidence indicating that the more complex model leads to a slightly better fit. However, the estimates of the within-area distribution obtained using the three models are very close. For illustration, scatterplots between those obtained with Models 1 and 3 are presented in Figure 3.

Table 1. Fitting results. Fixed effects models

Model	Age group	Estimated Coefficients	Deviance	Difference Deviance	Crit. value 5% Sig.
1) LA x Ethnicity Census 2011	-	b=1.007	4639.41 7358.58*	-	
2) LA x Ethnicity|Age Census 2011	0-4	b=0.990	1714.81	2) vs 1) 15.06	5.991
	5-15	b=0.963	2441.20		
	16 or more	b=1.010	3187.51		
3) LA x Ethnicity|Age Census 2011 & School Census	0-4	b=0.974; d=0.780	1703.21	3) vs 2) 56.51	7.815
	5-15	b=0.958; d=0.677	2414.87		
	16 or more	b=0.998; d=0.913	3168.93		

* Deviance of a Model 2 with b=1.007 in each age group.

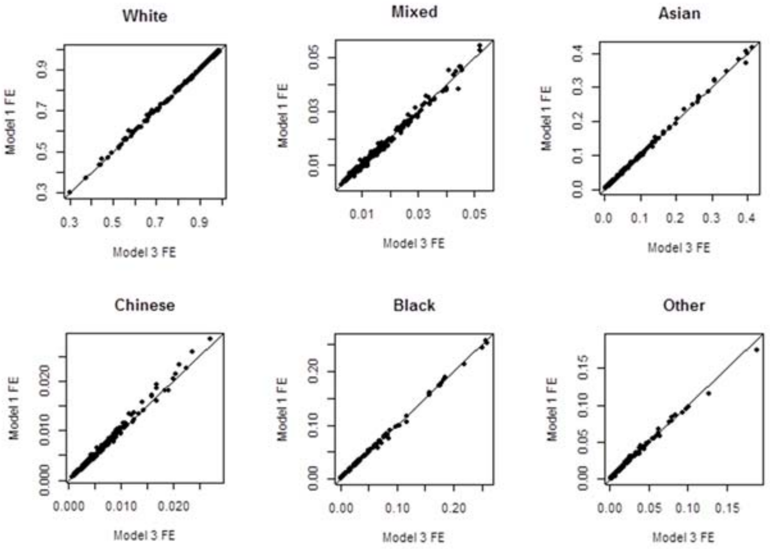

Figure 3. Estimates of the within-area distribution. Fixed effects GSPREE. Model 1 and Model 3. Line: Y=X

On the other hand, it is expected that if the sample sizes are big enough, an estimator based on a mixed effects model would be less biased than its fixed effects counterpart. For each of the models, we attempted to calculate the mixed effects version of the GSPREE proposed in Zhang and Chambers (2004) but it was impossible to achieve convergence in the estimation of the variance-covariance matrix of the random effects, possibly due to the generally low sampling fractions. As an alternative, we fitted a fully parameterised mixed effect GSPREE, with a log-link and a Poisson sampling distribution, similar to the one described by equation (3) but including cell-level independent random effects with category-specific variances, as an extension of Model 1. Only for three of the six categories positive estimates of the variance components were found. The estimates are 0.029, 0.014 and 0.101, for *White*, *Mixed* and *Asian* respectively. For the other categories, the corresponding variance component estimates were set to zero. The issue of negative variance component estimates for some but not all the categories will be discussed further in Section 5.

A set of scatterplots comparing the estimates obtained under the mixed effects version of the GSPREE estimator for Model 1 and the fixed effects version for Model 3 are presented in Figure 4. Differences in the estimated proportions are observed, especially for the categories Mixed, Asian and Chinese. Notice that even though for the last three categories the variance component estimate was zero, the two estimators do not coincide due to the IPF. Figure 4 does not suggest a bad performance of Model 3 in terms of bias, when compared to the mixed effect estimator.

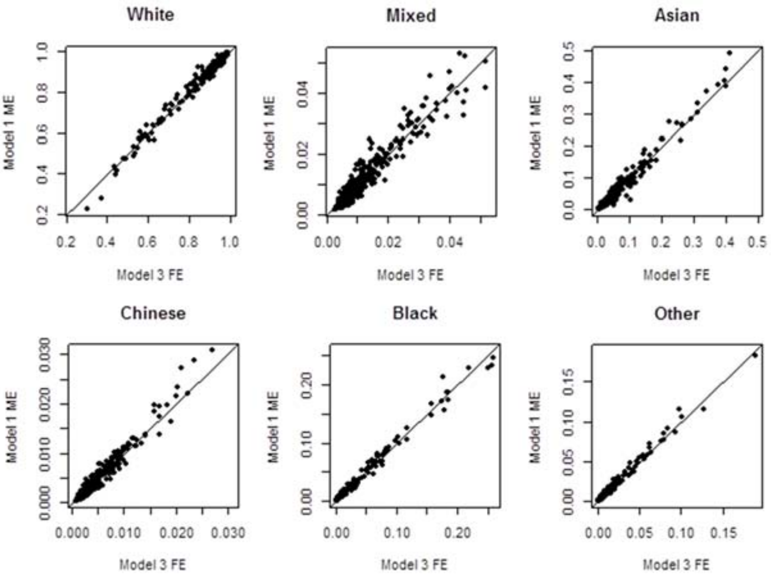

Figure 4. Estimates of the within-area distribution. Mixed effects GSPREE for Model 1, Fixed effects GSPREE for Model 3. Line: Y=X

4.1. Mean Square Error (MSE) evaluation

To assess the performance of the different estimators in terms of their Finite Population Mean Square Error (FP-MSE), a semi-parametric bootstrap approach was applied. The bootstrap samples were randomly generated from a plausible population composition, instead of randomly selected from a fixed synthetic population. Both approaches should perform similarly given that the implied sampling fractions of the APS are negligible but the former is considerably quicker. Two sampling designs were used: Multinomial, assuming the same observed sample size in each area as fixed, and Poisson sampling with random sample size. As counts by age are required to fit Models 2 and 3, independent samples were generated for each age group, and the aggregate of the three samples was used to fit the Fixed and Mixed effects estimators under Model 1.

The initial idea was to generate the population composition under a mixed effects model split by age. However, as mentioned before, it was impossible to obtain positive variance estimates for those models, and even in the case of Model 1, only three of the six categories have a positive variance component estimate. Using such variance components estimates could lead to an overly optimistic scenario for the GSPREE because of a lack of heterogeneity.

An alternative set of variance components was obtained from the two proxy tables, School Census 2012-2013 and Census 2011, by considering the School Census as a big sample from the true population in the age group 5-15 and using the methodology of the mixed effects GSPREE estimator. The estimated variance components are 0 for *White*, 0.02 for *Mixed*, 0.05 for *Asian*, 0.12 for *Chinese*, 0 for *Black* and 0.79 for *Other*. To allow for extra heterogeneity in all the categories, the two zero estimates were replaced by the minimum positive estimated value, 0.02. These estimates were used in all age groups, to generate the population composition from which the bootstrap samples are generated.

Despite the two zero estimates, we have found the set of variance components estimated using the two auxiliary sources more plausible than the one obtained from the sampling data under Model 1 in section 4.1, when taking into consideration the category specific heterogeneity observed in Figure 2. This could be seen as evidence against the performance of the mixed effects estimator presented in the previous section. It is possible that, even under Model 1, a synthetic estimator needs to be used given the small sample sizes in the cells of the survey composition.

The results in terms of FP-Bias and FP-MSE obtained under Poisson or Multinomial sampling were very similar, possibly due to the impact of the benchmarking on reducing the variability associated to the random area sample size in the case of the Poisson sampling. We will therefore omit one set of the results. The results for the Multinomial sampling are presented in Table 2 and Figures 5, 6 and 7.

Table 2. Average FP-Bias and Square root FP-MSE

		Ethnicity					
Measure	Model	White	Mixed	Asian	Chinese	Black	Other
Average FP-Bias	Model 1 FE	-0.00157	0.00013	0.00162	0.00028	-0.0001	-0.00036
	Model 2 FE	-0.0008	-0.00001	0.00119	0.00022	-0.0002	-0.0004
	Model 3 FE	-0.00156	0.00011	0.00164	0.00029	-0.00009	-0.0004
	Model 1 ME	-0.00158	0.00018	0.0016	0.00031	-0.00013	-0.00039
Average Square Root FP-MSE	Model 1 FE	0.00948	0.00195	0.00691	0.00158	0.00354	0.00717
	Model 2 FE	0.01177	0.00251	0.00895	0.00161	0.00418	0.00651
	Model 3 FE	0.00951	0.00189	0.007	0.00162	0.00357	0.00717
	Model 1 ME	0.00974	0.00187	0.00709	0.00167	0.00355	0.00716

Overall, there is no estimator that performs substantially better than the others, either in terms of FP-Bias or FP-MSE. Even though the average bias for each category is close to zero, according to Table 2, for specific areas there is bias in the estimation of the within-area distribution in all the fixed effects estimators, as it can be seen from Figure 5. The mixed effects estimator under Model 1 seems unable to correct this bias, given that the estimates with bigger biases are those for LAs with small sampling fractions. See Figure 6.

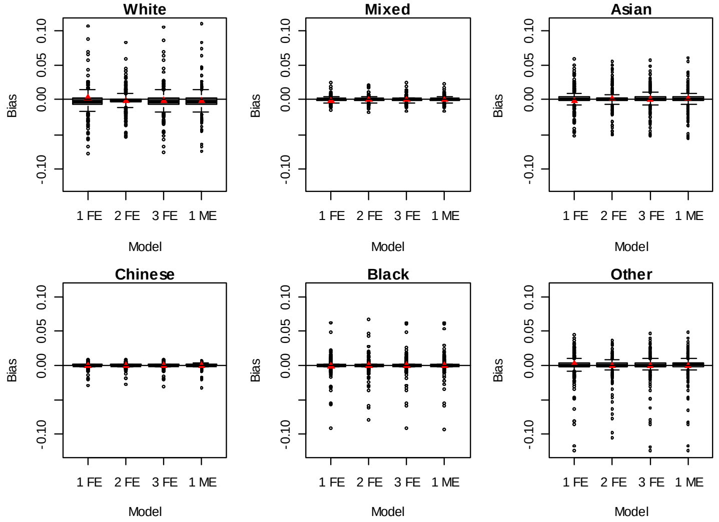

Figure 5. FP-Bias with respect to the simulated population composition. Red triangle: Mean

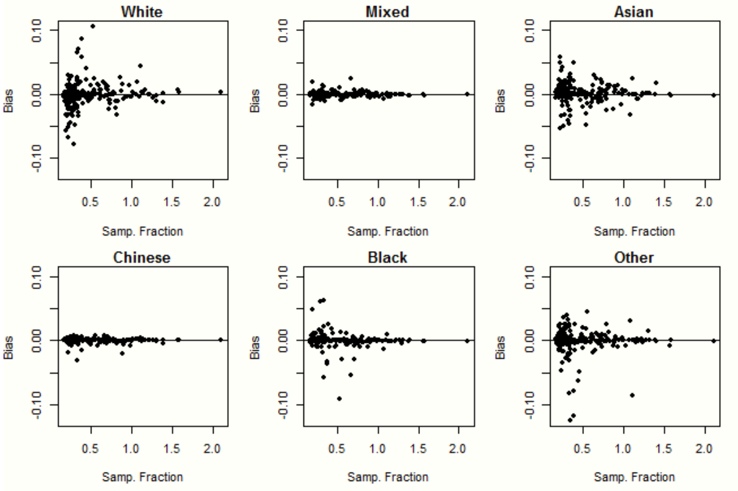

Figure 6. Implicit sampling fraction Vs. FP-Bias of the mixed effects estimator (Model 1)

5. Discussion

In this paper, we present a feasibility study to produce Small Area estimates of the within-area distribution of Ethnicity by LA in England, using the GSPREE. It is the first time this approach has been attempted for this type of problem in the UK. Unlike other demographic and socio-economic characteristics, Ethnicity is a variable for which there is no clear set of covariates identified in the literature, which could be used as a predictor. In fact, unless a proxy is involved, it seems difficult to expect good performance of a Small Area Estimator in this context. Structure Preserving Estimators can be used, given that proxy compositions can be obtained either from the last population census or from other sources, such as the School Census.

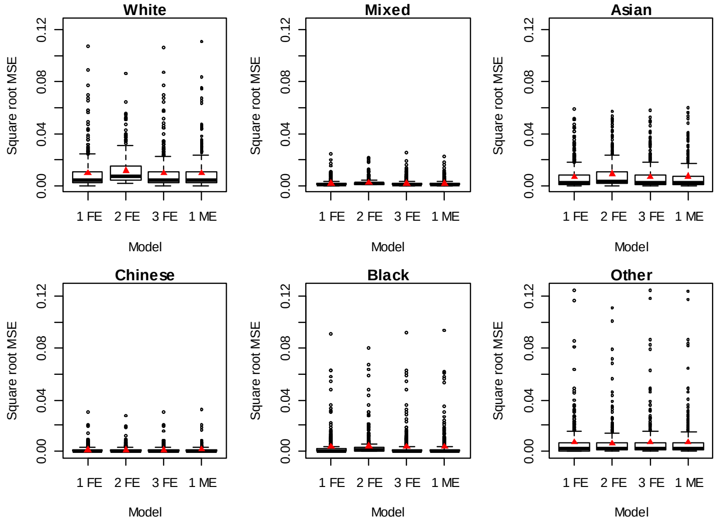

Figure 7. Square root FP-MSE with respect to the simulated population composition. Red triangle: Mean

In this work, we formulated three alternative models to produce the desired estimates with the GSPREE. However, in terms of Bias and FP-MSE, no substantial improvement was obtained by using more complex models or different sources of information. Moreover, given the small sample sizes available from the APS, synthetic estimates seem the only possible alternative in this case.

Notice that the lack of sample size to fit a mixed effects model is not a problem only of this application but rather one which all applications of SAE face sooner or later, if the aim is to produce estimates at increasingly lower levels of aggregation. In this sense, work to improve the synthetic predictor is of highest priority. Currently, we are working on a more flexible version of the fixed effects GSPREE and we expect to be able to evaluate it against the other estimators included in this paper, in the near future.

When it comes to mixed effects modelling, a particular problem we encountered with these data is that the variance component estimate can be negative for some but not all the categories, when the model allows for category-specific variance components. A possible remedy is to impose a common variance component. However, further study is needed in order to determine whether this or another random effects modelling strategy can be suitable.

Evaluation of the estimators in terms of their Bias and FP-MSE is also a topic for future work. The conclusions and quality of the evaluation is closely related to the plausibility of the characteristics of the artificial finite population, or as in our case, of the artificial population composition, from which the bootstrap samples are extracted. Additional work is still necessary in this area in order to formulate alternative scenarios that can be used to select a model, as well as to increase our knowledge on the performance of the proposed estimators.

REFERENCES

AGRESTI, A., (2013). Categorical Data Analysis. John Wiley & Sons.

BERG, E. J., FULLER,W. A., (2014). Small Area Prediction of Proportions with Applications to the Canadian Labour Force Survey. Journal of Survey Statistics and Methodology, 2 (3), 227–56.

CINCO, M., (2010). Intercensal Updating of Small Area Estimates. Unpublished PhD thesis. Massey University.

LIANG, K-Y., ZEGER, S. L., (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73, 13–22.

MOLINA, I., AYOUB S., LOMBARDIA, M. J., (2007). Small Area Estimates of Labour Force Participation under a Multinomial Logit Mixed Model. Journal of the Royal Statistical Society: Series A (Statistics in Society), 170 (4), 975–1000.

NOBLE, A., HASETT, S., ARNOLD, G., (2002). Small Area Estimation via Generalized Linear Models. Journal of Official Statistics, 18(1):45–68.

PFEFFERMANN, D., (2013). New Important Developments in Small Area Estimation. Statistical Science, 28 (1), 40–68.

PURCELL, N., KISH, L., (1980). Postcensal estimates for local areas (or domains). International Statistical Review, 48(1), 3–18.

RAO, J. N. K., (2003). Small Area Estimation. John Wiley & Sons.

SCEALY, J., (2010). Small Area Estimation Using a Multinomial Logit Mixed Model with Category Specific Random Effects. Research paper, Australian Bureau of Statistics.

UPTON, G., COOK, I., (2008). *A Dictionary of Statistics*. Oxford University Press.

ZHANG, L. C., CHAMBERS, R., (2004). Small area estimates for cross-classifications, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(2), 479–496.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 603–610

SAE TEACHING USING SIMULATIONS

Jan Pablo Burgard[1], Ralf Münnich[2]

ABSTRACT

The increasing interest in applying small area estimation methods urges the needs for training in small area estimation. To better understand the behaviour of small area estimators in practice, simulations are a feasible way for evaluating and teaching properties of the estimators of interest. By designing such simulation studies, students gain a deeper understanding of small area estimation methods. Thus, we encourage to use appropriate simulations as an additional interactive tool in teaching small area estimation methods.

Key words: small area estimation, teaching, simulations, design-based simulations, model-based simulations.

1. Challenges in Teaching SAE

Small area estimation (SAE) methods are becoming increasingly valuable for both methodologists and practitioners, and are used quite regularly in the production of official statistics. The last two decades have witnessed an explosion of small area estimation methods. However, the advances are mostly in the theoretical field, and practitioners still lack adequate knowledge of all the advancements in SAE methodology.

The classical way to present the benefits and drawbacks of SAE methods is using slides. Graphs and tables are used for illustration, and often simulation results are presented on the slides as well. From experience, however, for many students the understanding which estimator is preferably applicable is still lacking. These students, in order to obtain a good result in the exam, will memorize mainly the advantages and disadvantages of the respective methods. This is certainly not the didactic goal, and holds the further restraint that many of them are not able to transfer their knowledge to new methods developed later on.

We emphasize using simulations as an interactive tool to teach SAE methods. A large list of literature exists concerning the use of computers in statistical classes (McKenzie, 1992) and some papers focus directly on the use of simulations

[1] Trier University. E-mail: burgardj@uni-trier.de.
[2] Trier University. E-mail: muennich@uni-trier.de.

(Kalsbeek, 1996, Hesterberg, 1998, DelMas et al., 1999). Hesterberg (1998) describes simulations as follows:

> The basic idea in simulation is to emulate real life, where one collects a sample of random data (using a survey or an experiment), and summarizes the data graphically or numerically. In simulation one generates a sample of random data on the computer in a way that mimics a real problem and summarizes that sample in the same way. However, instead of doing this only once, one may do it many times, to investigate how much summaries vary.

In the context of statistical education Mills (2003) states that

> Regardless of how clearly a teacher explains a concept, students will understand the material only after they have constructed their own meaning for the new concepts, which may require restructuring and reorganizing new knowledge and linking it to prior or previous knowledge.

Further he points out that

> [...] meaning is acquired through a significant interaction with new knowledge.

An educational concept for teaching mathematics and statistics that addresses these aspects is *discovery learning* proposed, e.g., by Bruner (1961). The key idea is to provide students with materials needed to solve the imposed questions - rather than providing simply their solutions. However, as Mayer (2004) points out, an unguided form of discovery learning is not recommendable. Kirschner et al. (2006) state that the learners need guidance to reach a certain level of knowledge, from which point on they can increasingly learn from discovery. In an empirical evaluation of different teaching methods Alfieri et al. (2011) find that *Enhanced Discovery Learning* shows to have a positive effect on learning. In enhanced discovery learning, the teacher accompanies the discovery process by instructional guidance, or feedback or other merits. In our view, simulations provide a platform for such enhanced discovery learning with a built-in feedback system.

In the following section it is discussed how simulations can be used in the special context of SAE to support the students in the process of understanding the merits of the different methods at hand. In Section three, an example simulation used in graduate classes is provided. We conclude with a summary and outlook.

2. The Use of Simulations for Teaching SAE

In SAE, two major types of simulations can be considered, design-based and model-based simulations (for a more detailed discussion, see e.g. Burgard, 2013, and Münnich, 2014).

In model-based simulations random samples from a superpopulation model are drawn. The methods of interest are then applied to these random samples. This is an effective procedure to check particularly whether (a) under optimal conditions, that is when all model assumptions hold, a method yields the theoretically expected results and (b) a method is programmed correctly. Usually, it is far more sophisticated to derive a real world behaviour in the context of survey statistics. As Graham Kalton stated in Malay Ghosh's honorary symposium in 2014

> In case we want to apply small area methods in official statistics, we have to consider the sampling design.

In design-based simulations the random samples are drawn according to a sampling design from a fixed finite population. It is basically an attempt to reproduce the true survey process of interest. A major emphasis has to be laid on a realistic population that mimics all important characteristics of the real population. This realistic population could be for example an older version of the actual population. The design-based simulation then is useful for comparing different methods on their applicability in a certain survey context with regards to the sampling design.

Thus, when teaching SAE methods, model-based simulations are a good starting point to study the properties of SAE. However, for studying real world behaviour, the design-based simulation approach seems considerably more appropriate for applications, at least for official statistics.

As the field of SAE encompasses several statistical disciplines and applications, there are multiple decision criteria to acknowledge for when choosing appropriate methods. Some central but non-exhaustive aspects to consider are the classical statistical properties, user acceptance, as well as computational complexity and stability. Performing simulations in either way helps to understand advantages and disadvantages of the statistical methods given the relevant decision criteria, e.g. triple-goal (Shen and Louis, 1998), and further enables the students to evaluate new methods later on their own.

For most estimators in SAE, classical statistical properties are proven. These are generally based on asymptotic theory, regarding sample size, or the number of areas or domains. Both asymptotic arguments, however, have to be used carefully in SAE, as the typical setting is a small sample size and a finite number of areas (Pfeffermann, 2006). By varying the sample sizes within a simulation, the effect of small sample sizes or small number of small areas can be visualized. An example will be given in the next section.

An important hurdle is the acceptance of the published small area estimates by data users. This argument is specifically important in official statistics, where the users of the published data are not necessarily proficient in SAE. In practice, one major reservation against many small area estimators is that they are not design unbiased. However, as design unbiasedness and small variance of small area estimators are usually antagonists, the demand for design unbiasedness may better be dropped in favor of reducing the mse of the estimators. This can be visualized by using simulations.

SAE methods are often computationally very complex. Computation times may be prohibitive for too large data sets, and computational stability may depend severely on the data structure. Hence, the computability of many programs depends on the present sample. Using simulations, in general, a large set of different samples is provided and applied. Since many computer codes may fail in single samples, the simulation yields a realistic view on possible computational issues. Those *special* samples can be analysed into more detail which might lead to a reformulation of the estimator or an improvement of the computer program.

Additionally, in order to tackle in depth the before mentioned specific issues, simulations are a useful tool in the lecture to recapitulate the learned materials.

3. An Example for Using a Simulation in SAE Teaching

In general, when teaching SAE we start with the presentation of a new estimator and describe its statistical properties. Within the next step, students shall generate a superpopulation that fulfills all the assumptions of this estimator. The teacher accompanies the process of finding an appropriate superpopulation by asking supporting questions. By gradually deviating from the *optimal* superpoplation that fulfills all model assumptions of the estimator, the impact from deviations on the performance of an estimator can be observed.

Design-based estimation methods such as the direct estimator (Cochran, 2007, p. 21 et seqq.) rely on asymptotic arguments, and have good performance in large sample settings. Their performance, measured in terms of accuracy, is indirect proportional to the sample size. However, the sample size tends to be very small in SAE applications (Rao, 2003, p. 1). The following example simulation will tackle the following questions in this context. How do small sample sizes affect the outcome of direct estimators? Are there sample sizes under which we should prefer SAE methods to design-based methods? How much can we gain from using model-assisted and model-based estimation?

The students are asked to generate a superpopulation which shows the advantages of model-assisted and model-based estimation over the direct estimator without auxiliary variables. The discussion generally leads to the idea that the correlation between the dependent variable and the covariates, the ratio of between area variation and residual error, as well as the sample size will have an impact on the outcome of the different estimators.

The estimators of interest are the direct estimator without auxiliary information, the model-assisted direct estimator *GREG* (Särndal et al., 1992, §6.4), and the model-based Battese-Harter-Fuller estimator (BHF, Battese et al., 1988). From the viewpoint of official statistics, this may be seen as *from design towards model-based methods* (cf. Münnich et al., 2013). Holding the residual error constant, the superpopulation for a model-based simulation can be constructed with

- one dependent variable y as linear function of the realizations x of an arbitrary random variable X with

- normally distributed unit level error terms e with $\mathrm{E}(e)=0$ and $\mathrm{Var}(e)=\sigma_e^2$,
- and normally distributed area level error terms u with $\mathrm{E}(u)=0$ and $\mathrm{Var}(u)=\sigma_u^2$.

The resulting settings are as follows

- Setting 1: lower σ_u^2 lower cor(y,x)
- Setting 2: higher σ_u^2 lower cor(y,x)
- Setting 3: lower σ_u^2 higher cor(y,x)
- Setting 4: higher σ_u^2 higher cor(y,x)

By assuming higher and lower values for both, the correlation between y and x and for σ_u^2, the magnitude of the gain in efficiency of one estimator over the others can be visualized. As can be seen from Figure 1, the improvement of using the model-assisted as well as the model-based estimator over the direct estimator is the larger the higher the correlation between y and x. Additionally, the smaller the variance σ_u^2, and therefore the smaller the ratio $\frac{\sigma_u^2}{\sigma_e^2}$, the higher is the improvement over the direct estimator. Further, it becomes apparent that in the case of rather small sample sizes (n=4) the improvement of using the model-assisted and model-based estimators over the direct estimator without auxiliary covariates is larger than in the case of n=40. Especially the gain from using the BHF over using the GREG is more pronounced in the case of low sample sizes (n=4).

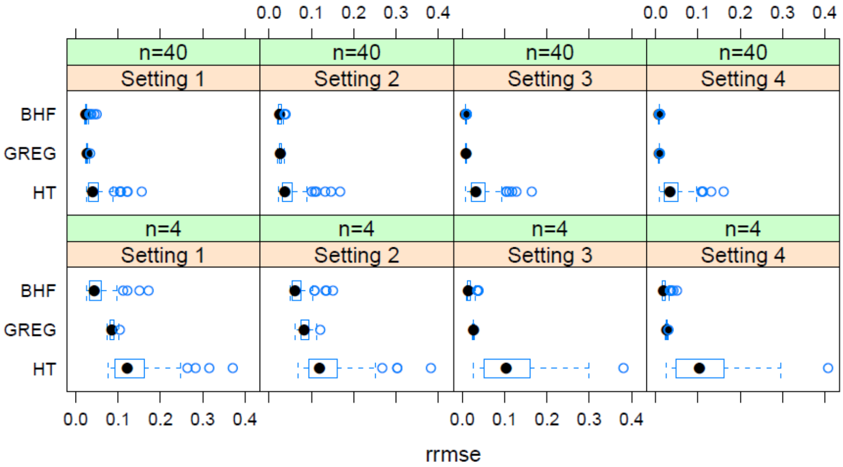

Figure 1. Rrmse of the estimators in the settings 1–4

Another perspective on the performance of estimators rather than looking at the rrmse, which is more convincing to many practitioners, is to look at the Monte-Carlo probability of lying within an acceptable interval. Such an acceptable interval can be defined as an interval in which the estimates should at least lie in. For instance, in Figure 2 an absolute distance of 1 from the true value is defined as acceptable. The Monte-Carlo probability of lying within the interval is then simply the rate of samples with successes within the Monte-Carlo simulation. The gain of using a model to not using auxiliary variables is immense. However, if sample size is larger, the gain from using the BHF over the GREG is not that pronounced as in the case of low sample sizes (n=4).

Certainly, in this context a considerable number of measures and their impact on the selection of adequate estimators can be investigated via simulations, which furnishes a better understanding of the entire methodology.

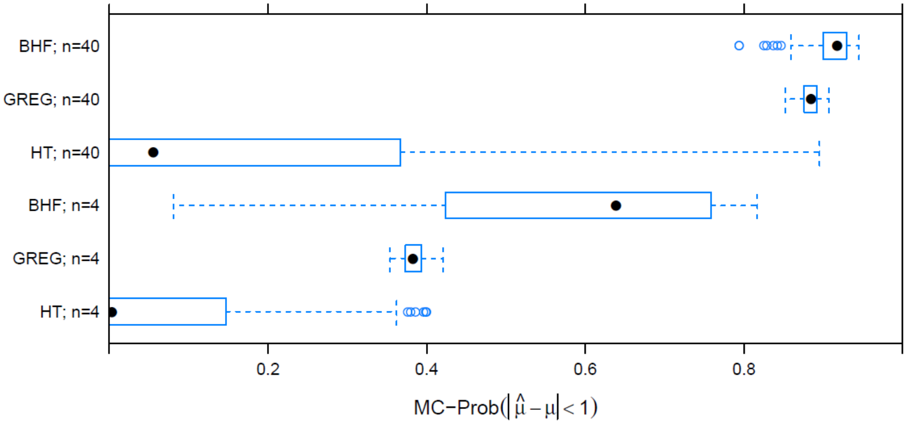

Figure 2. Monte-Carlo probability of lying within an acceptable interval in setting 3

4. Summary and Outlook

Teaching SAE methods covering both theory and applications is a challenging task. Students attending SAE classes rarely have a strong statistical education background with experience in applications. In this context we are convinced that the above presented approach of using simulation for teaching SAE methods is a very useful additional tool in teaching SAE. It provides a better and more sustainable understanding of applying and choosing appropriate SAE methods.

Acknowledgements

We thank Risto Lehtonen for inviting us to the Teaching SAE session at the SAE 2014 in Poznan. We also thank the audience for the feedback and the vivid discussion after the presentation and throughout the conference, as well as the editor and an anonymous reviewer for very valuable comments that helped to improve the clarity of this paper considerably.

REFERENCES

ALFIERI, L., BROOKS, P. J., ALDRICH, N. J., TENENBAUM, H. R., (2011). Does discovery- based instruction enhance learning? Journal of Educational Psychology, 103(1): 1.

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): 28–36.

BRUNER, J. S., (1961). The act of discovery. Harvard educational review.

BURGARD, J. P. (2013). Evaluation of Small Area Techniques for Applications in Official Statistics. PhD Dissertation, Universität Trier.

COCHRAN, W. G., (2007). Sampling Techniques. Wiley series in probability and mathematical statistics: Applied probability and statistics. John Wiley & Sons, New York, ISBN 9780471162407.

DELMAS, R. C., GARFIELD, J., CHANCE, B., (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. Journal of Statistics Education, 7(3).

HESTERBERG, T. C., (1998). Simulation and bootstrapping for teaching statistics. In American Statistical Association Proceedings of the Section on Statistical Education, pages 44–52.

KALSBEEK, W., (1996). The computer program called sample: A teaching tool to demonstrate some basic concepts of sampling (version 1.01). In American Statistical Association Proceedings of the Section on Statistical Education, Volume 103.

KIRSCHNER, P. A., SWELLER, J., CLARK, R. E., (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational psychologist, 41(2): 75–86.

MAYER, R. E., (2004). Should There Be a Three-Strikes Rule Against Pure Discovery Learning? The Case for Guided Methods of Instruction. American Psychologist, 59(1): 14–19, Jan. 2004. ISSN 0003066X. doi: 10.1037/0003-066X.59.1.14.

MCKENZIE J. D., (1992). Why aren't computers used more in our courses. In Proceedings of the Section on Statistical Education, pages 12–17. The Association.

MILLS, J. D., (2003). A theoretical framework for teaching statistics. Teaching Statistics, 25 (2): 56–58.

MÜNNICH, R., (2014). Small area applications: some remarks from a design-based view. SAE 2014-Conference on Small Area Estimation in Poznan, http://sae2014.ue.poznan.pl/presentations/SAE2014_Ralf_Munnich_c330a31c0a.pdf.

MÜNNICH, R. T., BURGARD, J. P., VOGT, M., (2013). Small Area-Statistik: Methoden und Anwendungen. AStA Wirtschafts- und Sozialstatistisches Archiv, 6(3): 149–191.

PFEFFERMANN, D., (2006). Invited discussion of paper by J. Jiang and P. Lahiri: Mixed model prediction and small area estimation. TEST, 15: 65–72, URL http:// eprints.soton.ac.uk/38527/.

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

SÄRNDAL, C. E., SWENSSON, B., WRETMAN J., (1992). Model Assisted Survey Sampling. Springer, New York.

SHEN, W., LOUIS, T., (1998). Triple-goal estimates in two-stage hierarchical models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(2): 455–471.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 611–630

SAE EDUCATION CHALLENGES TO ACADEMICS AND NSI[1]

Elżbieta Gołata[2]

ABSTRACT

The aim of the paper is to present some experiences in teaching Small Area Estimation (SAE). SAE education experiences and challenges are analysed from the academic side and from the NSI side. An attempt was undertaken to discuss SAE issues in a wider perspective of teaching statistics. In particular, the topics refer to Polish conditions, but they are presented against the background of selected international experiences and practices. Information comes from a special inquiry - a survey conducted among employees of statistical offices and academics from universities involved in SAE research. A further issue is inclusion of SAE in the EMOS project (European Master in Official Statistics). The survey is extended with information collected by monitoring of trainings and projects organized by the leading centres dealing with SAE. The results obtained are related to a similar survey within Eurostat project: ESSnet on Small Area Estimation, which was conducted in 2010. The study includes interest in learning and the need to implement SAE methodology, a range of subjects taught as well as a range of applications, forms of training, type of courses, software used and teaching methods. In particular, it intends to answer how strong the interest in small area estimation is, what the demand for practical and theoretical knowledge in the field is and what the recommendations for universities and statistical institutes are.

Key words: Small Area Estimation, statistical education.

[1] The paper is based on presentation prepared for the international conference on Small Area Estimation SAE 2014, which was held in Poznan (September 2014). The author wishes to thank T. Klimanek, who was the co-author of the *Survey on teaching, use and/or development of SAE methods*, for his help and inspiration.

[2] Poznań University of Economics. e-mail: elzbieta.golata@ue.poznan.pl.

1. Introduction

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write" (Samuel S. Wilks[3]). In the society of today, where every day we are inundated with a lot of information, the problem of statistical education is gaining greater and greater importance. The question of misunderstanding or misinterpretation of statistical data can be viewed in terms of the consequences. In addition to global issues, particular attention is paid to information at the local level, as it usually involves problems that are close to most of us.

The question of statistical education is extremely extensive and beyond the scope of this study. We restrict ourselves to one area of statistical research, which is Small Area Estimation (SAE). Conference on SAE held in Poznan in September 2014 was an occasion to raise a question on teaching SAE methods, its understanding, demand for specialists and problems connected with its applications. Justification for such limitation can be sought in the growing interest in this field of research and its importance. The history of modern survey sampling dates back to the mid of the 20th century when it grew considerably due to scientific developments, among others, in the work of Jerzy Neyman, Sir Ronald Fisher and Karl Pearson. Over time the range of topics investigated using survey methods has broadened enormously. We are witnessing an increasing demand for estimates at a lower level of geographic division than broad regions or countries. This is due to the growing importance of detailed information in policy-making, programs, allocation of government funds, in creating policy to ensure a balanced regional development. In response to this demand, in the mid-eighties of the last century, in May 1985 an international symposium was held in Ottawa. It was a joint initiative of Statistics Canada, The Laboratory for Research in Statistics and Probability of Carleton University and the Department of Mathematics and Statistics at the University of Montreal. This conference was certainly a very important event in the development of small area statistics. Many studies refer to it as a starting point towards the development of this field of statistical research. Small area estimation is particularly important for countries that undergo economic and social transition, in Central and Eastern European Countries, because of the decentralization process, development of free market economy, transfer of management to local authorities.

It should be stressed that the philosophy of small area statistics reflects holistic transformation of statistical research observed recently. First of all, it is the use of information from a variety of sources, including administrative registers (Wallgren and Wallgren 2007, 2014, Zhang 2012), estimation based not only on sample data, frequentist and model-depended approach (Ghosh and Rao 1994,

[3] This is a quote from the presidential address in 1951 of mathematical statistician Samuel S. Wilks to the American Statistical Association (JASA, Vol. 46, No. 253., pp. 1−18). Wilks was paraphrasing H. G. Wells from his book Mankind in the Making (full H.G. Wells quote is available at: http://osu.causeweb.org/cwis/SPT-FullRecord.php?P=FullRecord&ResourceId=1240).

Ghosh 2001, Rao 2003, Fuller 2009, Datta 2009, Burgard et al. 2014), model selection and checking (Datta at al. 2011, Pfefermann 2011), covariate selection, linear mixed models (Rao et al. 2014, Torabi 2015, Chambers et al. 2013), variance estimation (Maples, Bell and Huang 2009, Graf and Tillé 2014.), simulations, bootstrap (Burgard and Münnich 2014), calibration (Särndal 2007), benchmark (Hidiroglou and Smith 2005, Gosh at al. 2014), methods dealing with non-response (Särndal and Lundström. 2005, Longford 2005), data quality assessment (Wallgren and Wallgren, 2013) and many many others.

In particular, the aim of this paper was to present experiences and needs for teaching Small Area Estimation. Experiences in education and challenges were analysed, from the academic side and from the side of official statistics (National Statistical Institutes - NSIs). The study was aimed at answering the following questions:

- What is the experience in SAE methodology?
- How strong is the interest and demand for SAE methods, application, and teaching?
- What are the main problems in teaching SAE?
- What kind of risk should be considered when applying SAE?
- What are the most important sources of information on SAE developments?

The analysis allowed presenting differences in perception of particular SAE problems from the perspective of different institutions. When possible, changes observed in time were presented by referring to the results of the research conducted within Eurostat ESSnet project on SAE in 2010 (European Statistical System - ESS functions as a network). Answering formulated questions defined the structure of this paper. It starts with presenting the data and experiences of the surveyed institutions in SAE. The forms of activity in SAE, theoretical research and applications in the NSIs and at universities are compared. In the next section interest in teaching SAE is discussed by presenting needs for education and forms of teaching. The most important problems regarding teaching SAE are the subject of the third section. Risks and challenges in teaching SAE are discussed. It is followed by fourth section presenting an analysis of issues that attract special attention in a more general perspective of the European Statistical System, European Statistical Training Programme (ESTP) or European Master in Official Statistics (EMOS) project. Opinions on subjects that require special training as well as sources of information on SAE methodology are shown. Finally we summarize the results and draw some conclusions.

2. Experience in small area estimation of the surveyed institutions

Apart from assessing the progress in the development of Small Area Estimation methodology, this study addresses the issue of experiences and needs for teaching SAE. It was not an easy task, as knowledge in this area is not systematic. So in order to fulfil the aim of the study, to answer the formulated questions on SAE education, a special inquiry was conducted. Two questionnaires were prepared, one for institutions and the other for individuals. This distinction was introduced as not all questions were relevant to both institutes and researchers at universities (e.g. concerning the forms of teaching activities). But there were only small differences between both versions. Usually there were just slight changes in formulating questions. The questionnaires were sent via e-mails, but some of them were collected during the SAE 2014 conference in Poznan. The structure of the questionnaire responded to the objective of the study. It consisted of 11 questions regarding: experience in SAE, interest in teaching and demand for SAE methodology, problems and challenges for teaching and practical application as well as sources of information on SAE.

To obtain information a mailing list of people and institutions involved in SAE research and projects was used. This was the list prepared to disseminate information on SAE 2014 conference organized in Poznan. The mailing list contained addresses of statistical institutes of the European countries, especially those participating in Eurostat Projects on SAE, but also from other countries that had been previously involved in cooperation on this subject (it was an updated database used in a survey conducted within ESSnet on SAE project in 2010). In addition, the mailing list included addresses of scientists and researchers whose field of study is indirect estimation, who published articles on this subject, participated in earlier conferences organized by EWORSAE Council (European Working Group on Small Area Estimation) or ISI satellite conferences (e.g. The First Asian ISI Satellite Meeting on Small Area Estimation in Bangkok, September 1-4, 2013).

As a result of the survey 60 responses were obtained: 19 from statistical institutes and 41 from university researchers. Almost a half of the responses in the survey came from Poland (5 of statistical institutes and 22 from academics). In this way, the study reflects also perception of the role of SAE, as well as the possibility of its practical applications, and the demand for education in this field in Poland. It should be underlined that Poland is a country experiencing economic transformation and developing regional self-government, so there is a possibility of making comparisons with the results obtained for other countries. Thus, while discussing selected issues, they are accompanied by a reference to the situation observed in Poland.

The study obtained information from statistical institutes of different countries around the world. It should be emphasized that these are the countries with different systems of official statistics and using different methodology in statistical surveys. Among the countries whose statistical institutes participated in the survey, one can specify: Albania, Austria, Canada, Hong Kong, Japan, Kosovo, Latvia, Lithuania, Moldova, Romania, Slovakia, Suriname, Ukraine, USA. A comment might be added on the participation of the Statistical Institute of the Russian Federation in the survey. In this case a response was received that Rosstat does not conduct research on small geographic areas and for this reason cannot fill in the attached questionnaire.

Analysing responses of individuals, it can be noted that among 41 scientists who responded to the survey, about a half were Poles (22 persons). The remaining 19 participants of the study were researchers from countries like Australia, Canada, Finland, Italy, Japan, Norway, Spain, UK and USA.

In any case, the study was of no sample survey character. Nevertheless, the results are not only interesting, but also can be helpful in assessing the needs as well as identifying areas of further research. Additionally, the answers obtained often satisfied the 'saturation' condition used as a criterion for the number of in-depth interviews in sociological research.

Assessing respondents experience in SAE, one should remember that the mailing list contained people and institutions that were earlier involved in some kind of research on this subject. Apart from unrepresentativeness of the sample, it is worth noting that participation in seminars and conferences was also one of the forms of experience, as outlined in the next question. Thus, it can be expected that in each case at least minimal experience would be indicated. Meanwhile, despite a significant increase in experience, 13% of respondents admitted a complete lack of experience, as compared to the survey conducted four years earlier.

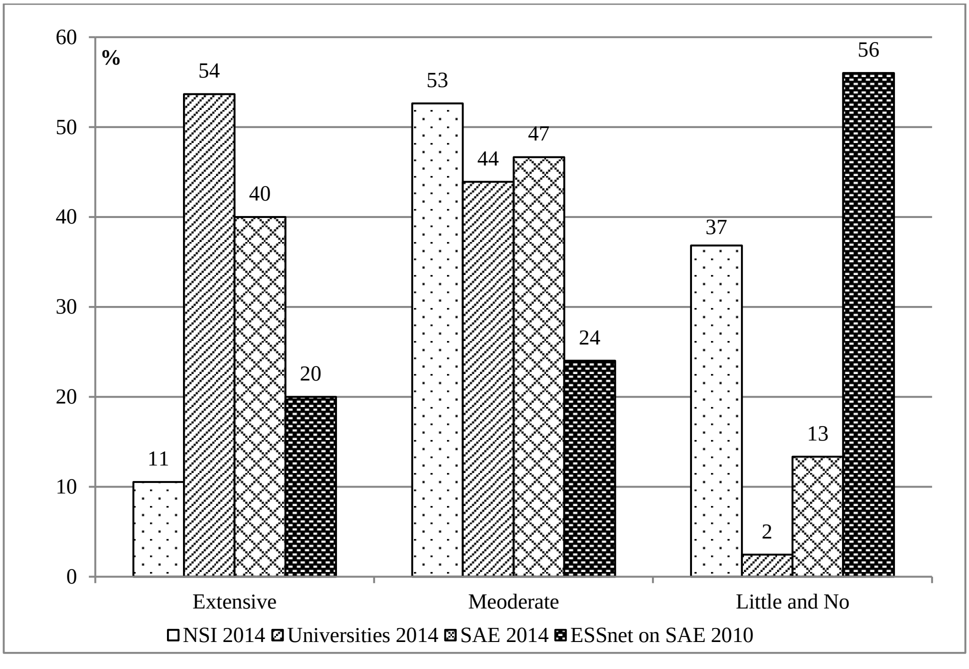

Figure 1. Experience in Small Area Estimation in surveyed institutions (%)

Source: Survey on teaching, use and/or development of SAE methods, July 2014, M. Szymkowiak, Report on the analysis of questionnaires used in WP 2, ESSnet on Small Area Estimation, 2010.

In 2010, the most of NSIs participating in the ESSnet survey admitted to have little experience when it comes to small area – 56% (see fig. 1). Four years later, the majority (64%) of the surveyed NSIs declared moderate (53%) or even extensive (11%) experience in SAE. Nevertheless, still 37% of them admitted to have little or no experience. There is no information to determine a similar trend among individual researchers at universities. Note, however, that among those surveyed, more than a half (54%) declared very extensive, with a further 44% of moderate experience in SAE. A result suggesting nearly 100% interest in the subject should not be surprising since the survey covered statisticians involved in this area of research and their opinion is the subject of analysis below.

It is interesting to compare forms of experience of the NSIs and universities (see tab.1). Statistical Institutes pointed primarily to practical applications as a basic form of activity in SAE (37%). However, among the academics scientific work and participation in seminars and conferences was the dominant form of experience (to the same extent, 51% of responses).

Table 1. Experience in Small Area Estimation by form of activity (%)

Form of experience in SAE	**NSIs**	**Univer sities**	**Total**
Theoretical (e.g. scientific research, literature studies)	16	39	32
Scientific research including:	16	51	40
Experimental research (e.g. simulations carried out on unreal data)	5	32	23
Development research (e.g. comparative analyses conducted on real data, assessing quality of the estimates, testing different estimators, models, etc.)	16	49	38
Practical applications (published estimates)	37	41	40
Scientific conferences, seminars and discussions	21	51	42
Teaching	16	27	23
Participation of NSI/your institution employees in lectures, seminars, courses	16	34	28
Joint projects (Eurostat etc.)	11	32	25
None	0	2	2

Notice: Percentage of the number of indications among a specified group of respondents. There was a possibility of choosing more than one issue and therefore the percentages do not sum to 100%.

Source: Survey on teaching, use and/or development of SAE methods, July 2014.

For statisticians from NSIs, the attendance in conferences and seminars was the second most common form of experience (21%). Other types of activity may be divided into two groups. The first is theoretical, scientific, and development research (16%). The second group is associated with teaching SAE: organizing and conducting courses, as well as participation in training and lectures (16%).

On the other hand, as for the experience indicated by academic teachers, development (49%) and theoretical (39%) research as well as practical applications (41%) should be emphasized, in addition to the already mentioned scientific work. Participation in lectures and joint research projects (e.g. Eurostat) could also be mentioned among important forms of SAE activities (34-32%). It is worth noting that teaching SAE (27%) was not the most common form of activity among academics.

Summarizing the results obtained for all the respondents, seminars and conferences (42%), as a forum for exchange of knowledge and experiences, are the most common form of activity as concerns SAE. In terms of frequency, the second and certainly no less important form is scientific research and practical applications (40%).

3. Interest in teaching and demand for sae methodology

The purpose of the study was, inter alia, a practical review feedback on the training needs of SAE. Among statisticians, there is a common belief in the need for education of SAE. This seems understandable among people who are engaged in this field of statistical research. The need for practical research is raised in almost in every study on this subject (see Platek et al.1987, Rao 2003). Clearly due to this belief the need for training seems to be understandable. However, it is also noted that statisticians and researchers who do not deal with indirect estimation often express their willingness to treat it as a cure-all for any shortcomings on the availability of data and estimation problems.

The results from the survey show that researchers who were themselves involved in SAE considered the need for NSI staff with methodological knowledge on indirect estimation as great (42%) and moderate (41%, see fig. 2). None of respondents participating in the survey considered education in this field as unnecessary. Comparing the opinions expressed in this regard by NSIs and Universities, we note that the frequency of indicating very big demand was twice as high for the official statistics (26%), compared to 13% among academics. This resulted in more often expressed moderate opinions on the educational needs among academics than official statisticians.

The results presented above seem to be very reasonable. They reflect strong demand for small domain estimates, articulated by official statistics and other recipients of their products. On the other hand, the results obtained for academics, confirm the importance of and the need for training, but keep a greater distance.

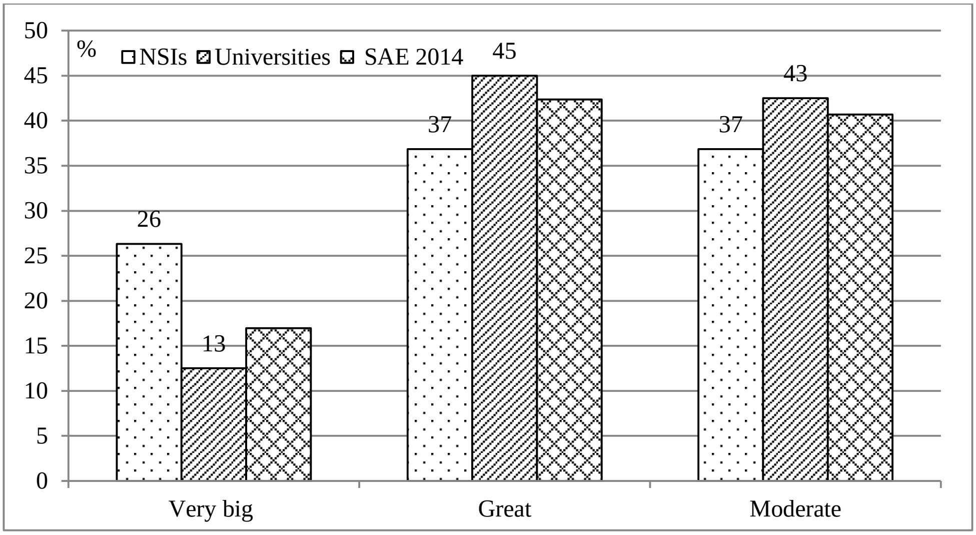

Figure 2. Need for education of NSI staff in SAE methodology in the opinion of NSI and Universities (%)

Notice: Percentage of the number of indications among a specified group of respondents. There was a possibility of choosing more than one issue and therefore the percentages do not sum to 100%.

Source: Survey on teaching, use and/or development of SAE methods, July 2014.

The results of the study do not allow us to say with certainty what is the demand for experts in SAE. Analysing the opinions expressed in the survey, it should be noted that the majority (58%) of respondents from NSIs agreed with the statement that SAE is one of the most desirable area of study. Additionally, almost all NSIs stressed that they do not employ specialists in SAE.

Analysing the demand for statisticians, specialists in specific methods of research, a question about the most important issues that require increased knowledge of NSI's staff, was included in the questionnaire. It was addressed to both statisticians from universities and from statistical institutes, allowing a choice of several of the listed options. The results obtained fully meet the requirements of the era of Modern Information and Communication Technology (ICT). The majority of respondents (65%) indicated training in statistical software as the most important issue: 745% of statisticians working in NSIs and 61% from universities (fig. 3).

It is also worth noting that in the students' opinion (e.g. of those who study Computer Science and Econometrics at the Poznan University of Economics, see PKA, 2014), modules for teaching specialized software are of special interest. Therefore, in order to attract a specific field of study, an educational offer often found objects taking into account that demand (e.g. Practical Data Science with R, Data Mining with SAS Enterprise Miner, The statistical analysis of market research with IBM SPSS Visualization and reporting of statistical data R / SAS).

An increasingly common practice is also to offer students a choice of training modules that allow them to obtain additional certificates honoured in the labour market (e.g. SAS Global Certification program, SPSS certificate of Expert Technology, SAP certificate).

Table 2. Most important issues that require increased knowledge of NSI's staff in the opinion of NSIs and Universities

Methodological issue	**NSIs**	**Universities**	**Total**
Sampling	9	14	23
Calibration	8	8	16
Spatial analysis, e.g. GIS	13	19	32
Big Data	8	13	21
Software (SAS, R, SPSS, etc.)	14	25	39
The choice should be left to individuals to be compatible with their interests	5	10	15
There is no need for NSI staff to train in any area	0	0	0
Other, what?	0	2	2
Number of Respondents	**19**	**41**	**60**

Source: Survey on teaching, use and/or development of SAE methods, July 2014.

Coming back to the most important problems that need to broaden the scope of teaching, the second most common group, as in the opinion of the NSIs (68%) and academics (46%), are methods of spatial analysis (53% of all respondents, see fig.3 and tab. 2). These are very important issues that are becoming more and more popular, among other things, due to the development of Geographical Information Systems (GIS). But it is worth noting also a direct relationship of spatial analysis methods to SAE, which is also reflected by the most common use of indirect estimation. This is also visible in the most commonly used term for this area of research: Small Area Estimation instead of Small Domain Estimation.

Sampling is only third in the 'ranking' of the most desirable skills. The survey does not mention explicitly SAE as an important problem that requires a broader education. However, respondents had the opportunity of individual declarations. Only two out of 60 surveyed took advantage of this opportunity. Among these were indications of modelling, especially for large, complex data sets and statistical inference. None of the respondents pointed directly at SAE. Perhaps this is a result of a mature approach, indicating particular problems that are of great importance in improving the quality of indirect estimation, or simply SAE is considered as an integral part of the sampling methodology.

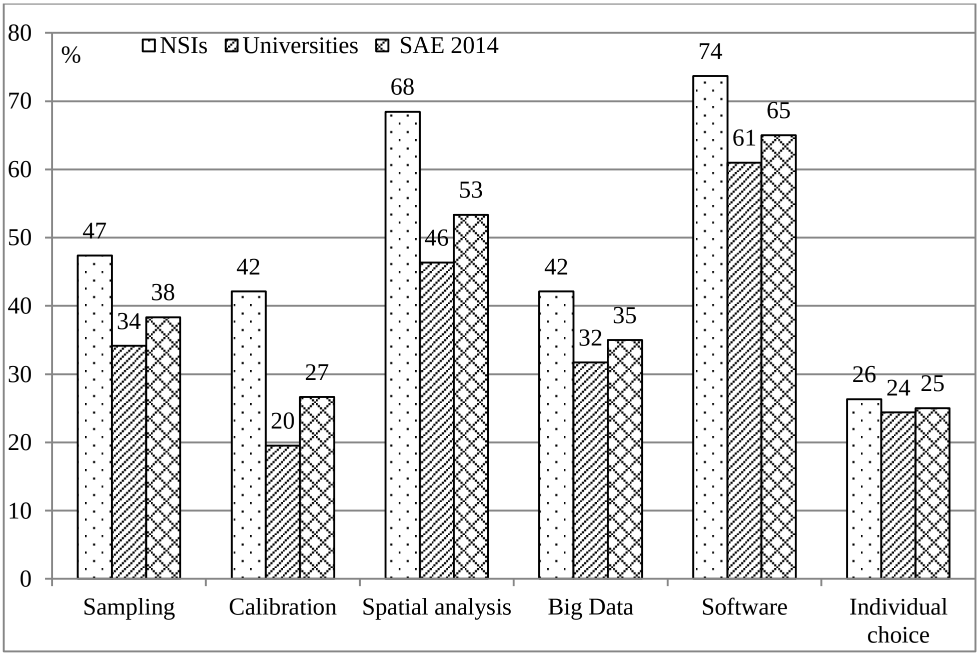

Figure 3. Most important issues that require increased knowledge of NSI's staff in the opinion of NSIs and Universities (%)

Notice: Percentage of the number of indications among a specified group of respondents. There was a possibility of choosing more than one issue and therefore the percentages do not sum to 100%.

Source: Survey on teaching, use and/or development of SAE methods, July 2014.

The above considerations should be complemented by another comment. The survey results indicated a much greater demand for skills in calibration reported by NSI staff (42% of responses) compared to 20% of such opinion among academics. It is understood that the demand for the ability to calibrate is much more appreciated by statistical institutes than among academics. One might be surprised by 10 percentage points difference in the number of indications of the importance of education in the area of Big Data. It was observed that 42% of responses on the significance of this issue was among statisticians from NSIs compared to 32% among academics. Big Data still raises lively discussion, many doubts and controversies.

4. Experience in teaching SAE, problems and challenges

Information on SAE education was obtained from surveyed statisticians who worked at universities. Academics were asked if their universities were teaching SAE. It turned out that SAE teaching experience was not large. Among surveyed respondents about 44% declared teaching SAE at their universities. But SAE education was rather not a part of regular lectures. Regular SAE courses are taught only at few universities, these are University of Southampton, Pisa, Trier, Helsinki, and University of Maryland and University of Michigan (Lehtonen 2014). Often basic information about SAE methodology was presented during other modules and lectures. In most cases these were seminars during master or post-doctoral studies.

Trainings and workshops organized in connection with scientific conferences were another often practiced forms of education in SAE. A similar form was taken by SAE dissemination among the employees of the institutions concerned. Trainings and workshops organized for special needs of other institutions, such as statistical offices were also listed among other forms of SAE education.

Trainings were one of the most common forms of education in SAE. They were organized both by universities, as well as NSIs. There were at least a few centres offering trainings and courses on SAE. The following list is not complete, but one could mention here University in Southampton with a 10 year long tradition of Southampton Statistical Sciences Research Institute, Universities in Helsinki, Pisa, Trier, Pompeu Fabra University in Barcelona, Statistics Finland and Istat (Italian National Institute for Statistics). Particularly noteworthy is the European Statistical Training Programme (ESTP) offered by European Statistical System (ESS). ESS is the partnership between Eurostat (with a leading role) and NSIs that are responsible for the development, production and dissemination of statistics. The purpose of the ESTP is to provide the opportunity to participate in international training courses at postgraduate level and other learning opportunities.

Another important ESS initiative was the establishment of the European Master in Official Statistics (EMOS) project. EMOS is an infrastructure project aimed at developing a programme for training and education in official statistics within existing master programmes at European universities. Increasing demand for quality information is widely recognized implying a strong investment in training of statistics. Programs like EMOS can be a part of the answer to these needs (Sorvillo, 2014). EMOS is planned to provide certified training in methodologies, statistical surveys, statistical production, analysis and statistical law and should be offered by a network of NSIs and Universities. Among others, *Survey methodology* and *Small Area Estimation* are in the list of elective courses.

Statistical software used in SAE was already discussed, but it still certain focus. ICT and software used in indirect estimation, modelling, simulations, bootstrap, etc. is of significant importance to the development of SAE. Therefore, compatibility of software used in NSIs and at universities for educational purposes might be desirable.

In the SAE2014 survey, 13 out of 19 NSIs expected knowledge of special statistical software, pointing SAS as the most preferable. However, the most popular software used in SAE, both in teaching and scientific research, was R. It was indicated almost twice as often as SAS (51% of respondents versus 28%, see fig.4). SAS software was used mostly in the NSIs, but becomes less popular among academics. Other software, like Bugs, WinBugs or SAS were rarely indicated. There were also institutions, like, e.g. Statistics Canada, which declared to develop their own, very flexible software for SAE.

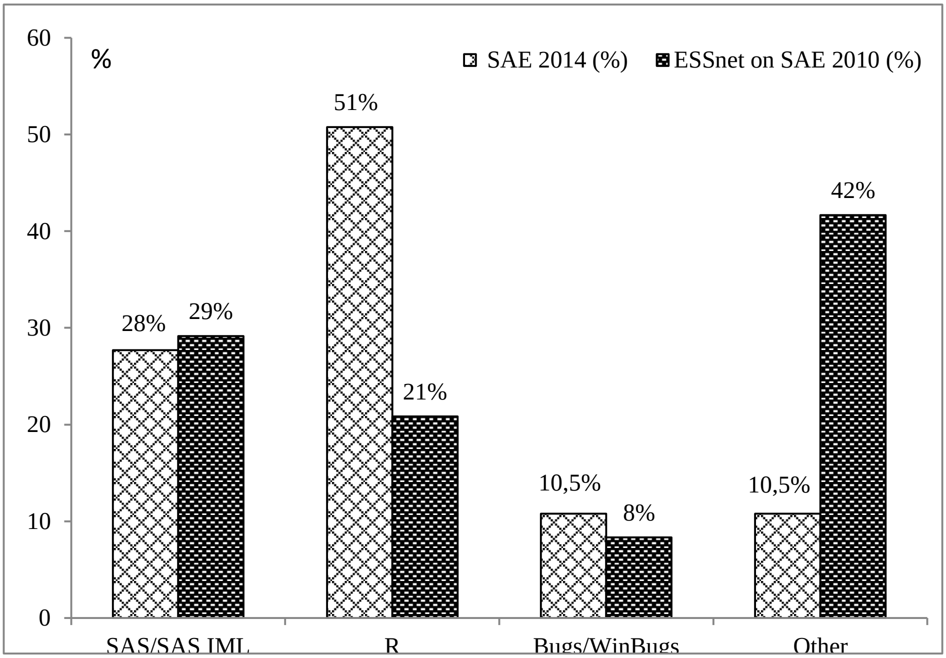

Figure 4. Software used in Small Area Estimation (%)

Source: Survey on teaching, use and/or development of SAE methods, July 2014, M. Szymkowiak, Report on the analysis of questionnaires used in WP 2, ESSnet on Small Area Estimation, 2010.

Comparing trends observed in time, once again a huge increase in popularity of R should be stressed. A good illustration would be the fact that in 2014 80% of all responses were indicating R, while in 2010 it was 21%. In this sense, the implementation of SAS is also gaining popularity, as in 2010 its score was 29% of responses in comparison to 44% in 2014. The increasing use of R and SAS program was associated with a clear reduction of interest in other software.

In the discussion so far the need for teaching of statistics was underlined, in particular small area statistics. But the results obtained show relatively little interest, with a very small number of universities, where SAE teaching programs were implemented. Recognizing these difficulties, a request to identify problems and challenges in teaching SAE was formulated. Subsequently a question about problems and risk in applying SAE was asked.

At first, the problems in teaching SAE will be discussed. The question on problems was of an open form. Answers collected here were divided into four groups. The first group contained statements indicating a lack of adequate preparation of students. The second one consisted of indications showing low awareness of not only the demand for specialists in SAE, but of ignorance of the existence of such a field of statistical research in general. In the third group demands on improving the attractiveness of classes of small area statistics were placed. The last group included suggestions to use the uniqueness of SAE for the dissemination of knowledge. It seemed to be worth listing to some of the opinions expressed in the survey, as it was thought of as a tool to share different views on the problem and how to cope with it.

1. Inadequate preparation among students
   - Students are not sufficiently prepared
   - Made it more understandable for students
   - Lack of sufficient knowledge of mathematical statistics
   - Problem with understanding basic methods in survey sampling, not only SAE
   - Some potential students may not be very familiar with statistical modelling or Bayesian inference
   - Lack of knowledge of statistical software
2. Low awareness of SAE as the field of statistical research
   - Low awareness of the need for knowledge and the development of SAE methods
   - Little popularity of the field
   - Small number of experts
   - Lack of textbooks in Polish
   - Use basic sampling course to generate interest of students in SAE
   - Reluctance of students to quantitative subjects
   - Recognition of the purpose of SAE
3. The attractiveness of classes
   - Careful preparation of teaching materials which allows full interaction between participants and the lecturer
   - Case studies based on actual research and applications
   - Incorporation of recent deliverables in teaching
   - More extensive use of multivariate data analysis in SAE, especially in the selection of auxiliary variables
   - Teaching students to build appropriate statistical models for use in SAE
   - Linking SAE and GIS
4. Unique challenges
   - Process approach to teach small area statistics - use of different methods and data sources
   - Understanding the capabilities and limitations of SAE
   - Teaching students to realise the difference between practical and theoretical approaches to SAE

– Make people understand the differences between the different approaches for inference (design-based, model-assisted, model-based).

The question about problems with teaching SAE somehow forces the question of problems and risks associated with the use of SAE methods. If they were commonly known and used in the majority of research and analysis, it can be assumed that they would be more familiar. Knowledge of these methods would identify the demand for specialists and experts in the field. So, it is worth to consider problems and risks arising from the use of SAE methodology. In addition, SAE methods require highly advanced knowledge, it is rather impossible to teach them at primary level. They are difficult and, as pointed above, student's knowledge is often insufficient to pass a basic course in survey sampling.

The respondents participating in the survey emphasized mainly bias of indirect estimators (62% of respondents, see tab. 3) and model-based approach (50%) as primary risks of practical application of SAE methods. More often the bias problem was pointed by academics than statisticians from the NSIs. Statisticians in official statistics institutions most frequently emphasized difficulty in variance estimation (53%), what in comparison to opinion of academics and all respondents was in fourth place. This relationship may seem a bit surprising, as NSIs put great emphasis on quality of the estimates (including not only accuracy but also unbiasedness).

Table 3. Problems and risks of applying SAE in official statistics (%)

Problems and risks	**NSIs**	**Universities**	**Total**
Small sample size	42	39	40
Bias of indirect estimators	47	68	62
Model dependent approach	42	54	50
Difficulty in estimating the variance	53	32	38
There are no risks	5	2	3
Other	11	2	5

Notice: Percentage of the number of indications among a specified group of respondents. There was a possibility of choosing more than one issue and therefore the percentages do not sum to 100%.

Source: Survey on teaching, use and/or development of SAE methods, July 2014.

Among statisticians from universities, a small sample size (39%) and the difficulty in estimating variance (32%) were mentioned as less important.

The frequency distribution indicating importance of such problems as sample size, bias, model dependent approach and variance estimation, was more uniform among statisticians from NSIs. 11% of responses from NSIs indicated “Other” issues, not specified in the study of problems in application of SAE. These results clearly show the difference between NSIs and universities. This is probably associated with the difference of the statutory tasks of both institutions: providing estimates and education, while the development of science is the common denominator.

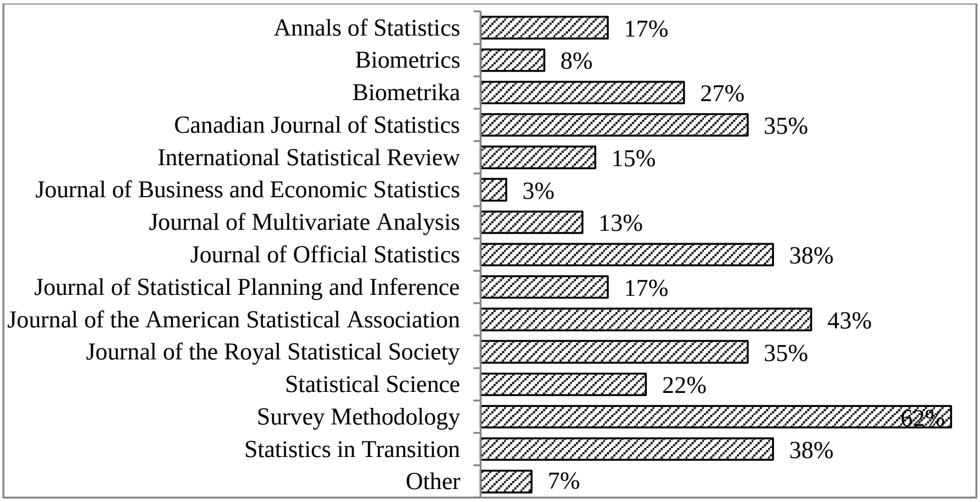

Figure 5. Main sources of information on SAE methodology and applications (%)

Source: Survey on teaching, use and/or development of SAE methods, July 2014.

Development of SAE methodology needs deepening and dissemination of knowledge and information about proposals of new methods, results of the research conducted and simulation. Statistical literature is very extensive. As regards SAE the most specialized scientific journals are: Survey Methodology (62%), Journal of the American Statistical Association (43%), Journal of Official Statistics (38%), Statistics in Transition (38%), Canadian Journal of Statistics (35%), Journal of the Royal Statistical Society (35%), Biometrika (27%) (see Fig. 5).

SAE is a relatively new area of statistical research. People dealing with these topics mostly know each other (if not personally, then through papers and via Internet). R. Lehtonen (2014) defines this as SAE "ecosystem". Its important part includes conferences and seminars organised by international organisations like International Statistical Institute or International Association of Survey Statisticians. Important contributions were made by conferences in Jyväskylä, Pisa, Elche, Trier, organized within the framework of the European platform called EWORSAE, that is European Working Group on Small Area Estimation which was founded in 2007 in Pisa.

Other components of SAE “ecosystem” mentioned by R. Lehtonen are those of U.S. SAIPE Program: Small Area Income and Poverty Estimates (Kalton and Citro 2001). On the European side there are European Union Framework Programmes for Research and Technological Development (FP) research projects conducted by Universities and NSIs. Several projects could be mentioned here: EURAREA, AMELI, SAMPLE. Their output is huge, not only through the development of knowledge, but also its dissemination and the introduction of specific forms of cooperation between the different centres. European Statistical System initiated also another program aimed at development of a framework for the production of small area estimates for ESS social survey. Essential elements of SAE "ecosystem" are of course books (Rao, 2003, Longford 2005, Fuller 2009), manuals, scientific and working papers, presentations and research reports.

5. Conclusion

The results of the *survey on teaching, use and development of SAE methods,* other available information from previous studies and the Internet allow for concluding that there is a growing awareness of SAE methodology. More than a half of the surveyed NSIs declared to have moderate experience in SAE.

However, this experience in SAE was mainly participation in seminars and conferences, but also scientific research was often mentioned. Therefore, great needs for education of NSI staff in SAE methodology was expressed by over 60% of NSIs. But it was not SAE methodology that was indicated as the field that requires increased knowledge of NSI staff. In view of the survey, the most important issues that need to broaden the scope of teaching were statistical software (SAS, R, SPSS, etc.) and methods of spatial analysis (GIS). As regards the software used in Small Area Estimation, R was the most popular in scientific research and SAS was the software most often used in the NSIs.

Referring to the main problems and challenges in teaching SAE, many opinions were expressed on poor preparation of students for advanced topics. Low awareness of SAE as the field of statistical research was also underlined. Suggestions could be found to use basic sampling course to generate interest of students in SAE. It was proposed to increase attractiveness of classes by introducing case studies based on actual research, practical implementation and incorporation of recent deliverables in the teaching process. Among unique challenges of SAE applications, the importance of understanding the capabilities and limitations was recognized as essential.

Problems in teaching are not independent form problems of SAE application. An analysis in this field pointed out first of all to bias of indirect estimators. However, for statisticians from NSIs, the main problem was difficulty in variance estimation. But they also stressed many other problems unspecified directly in the study. The main source of information on SAE methodology and applications is undoubtedly *Survey Methodology*. Among other scientific journals one should emphasize the importance of *Statistics in Transition*, particularly as a journal with a large audience among statisticians in Central and Eastern Europe.

Being aware of the limitations of the analysed survey, it is our hope that the results presented will help to bring together opinions of scientists from universities and practitioners working in National Statistical Offices and other institutions. And allow to confront and compare they own ideas and experiences with those expressed by colleagues representing the academic community, official statistics, research centres as well as other institutions involved in developing and applying small area estimation methods.

REFERENCES

MAPLES, J. J., BELL, W. R., HUANG, E. T., (2009). Small Area Variance Modelling with Application to County Poverty Estimates from the American Community Survey Statistical Research Division, U.S. Census Bureau, Washington, DC.

BURGARD, J. P., MÜNNICH, R., (2014). SAE teaching using simulations, presentation during International Conference on Small Area Estimation SAE Poznan 2014.

BURGARD, J. P., MÜNNICH, R., ZIMMERMANN, T., (2014). The Impact of Sampling Designs on Small Area Estimates for Business Data, Journal of Official Statistics, Volume 30, Issue 4.

CHAMBERS, R., CHANDRA H., SALVATI, N., TZAVIDIS, N., (2013). Outlier robust small area estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Volume 76, Issue 1, pages 47–69, January 2014.

CHANCE, B., (2002). Components of Statistical Thinking and Implications for Instruction and Assessment. Journal of Statistics Education Volume 10, Number 3 (2002).

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

DATTA, G. S., HALL, P., MANDAL, A., (2011). Model selection by testing for the presence of small-area effects, and applications to area-level data. Journal of the American Statistical Association, 106, 361−374.

FULLER, W. A., (2009). Sampling Statistics, Hoboken, New Jersey: John Wiley & Sons.

GHOSH, M., (2001). Model-Dependent Small Area Estimation – Theory and Practice, in: Lectures Notes on Estimation for Population Domains and Small Areas, eds. R. Lehtonen, K. Djerf, „Reviews" no. 5, Statistics Finland, University of Jyväskylä.

GHOSH, M., RAO, J. N. K., (1994). Small Area Estimation: An Appraisal, „Statistical Science", Vol. 9, No. 1.

GHOSH, M., KUBOKAWA, T., KAWAKUBO, Y., (2014). Benchmarked Empirical Bayes Estimators for Multiplicative Area Level Models. presentation during International Conference on Small Area Estimation SAE Poznan 2014.

GRAF, E. TILLÉ, Y., (2014). Variance Estimation Using Linearization for Poverty and Social Exclusion Indicators. Survey Methodology. June 2014, Vol. 40, No. 1.

HIDIROGLOU, M. A., SMITH, P., (2005). Benchmarking through calibration of weights for microdata. Working Papers and Studies, European Communities, Eurostat, Luxembourg.

KALTON, G., CITRO, C. F., (2001). Small-Area Estimates of School-Age Children in Poverty. Division of Behavioral and Social Sciences and Education, Commission on Behavioral and Social Sciences and Education, Committee on National Statistics.

LEHTONEN, R., (2014). Experiences and challenges in teaching Small Area Estimation, presentation during International Conference on Small Area Estimation SAE Poznan 2014.

LONGFORD, N. T., (2005). Missing Data and Small-Area Estimation, Springer.

PFEFFERMANN, D., (2011). Modelling of complex survey data: why is it a problem? How should we approach it? Survey Methodology, 37, (2), 115−136.

PFEFFERMANN, D., (2013). New important developments in small area estimation. Statistical Science, 28, (1), 40−68.

PLATEK, R., RAO, J. N. K., SÄRNDAL, C. E., SINGH, M. P., (1987). Small Area Statistics. An International Symposium. John Wiley & Sons. Ltd.

PKA 2014. The report on the program assessment held on 17-18 May 2014 for the field of study: computer science and econometrics conducted within the area of social sciences (first- and second-degree) at the Faculty of Informatics and Electronic Economy, Poznan University of Economics, State Accreditation Commission),

http://ue.poznan.pl/data/upload/articles/20141031/ce25bb287995611010/011-4-raport-pka-2014.pdf.

RAO, J. N. K., (2003). Small Area Estimation, John Wiley & Sons. Ltd.

RAO, J. N. K., SINHA, S. K., DUMITRESCU, L., (2014). Robust small area estimation under semi-parametric mixed models. Canadian Journal of Statistics, 42(1), 126–141.

SÄRNDAL, C-E. LUNDSTRÖM S., (2005). Estimation in Surveys with Nonresponse, John Wiley & Sons, Ltd.

SÄRNDAL, C-E., (2007). The Calibration Approach in Survey Theory and Practice, Survey Methodology, Vol. 33, No. 2, 99–119.

SORVILLO, M. P., (2014). EMOS as a new tool for training professionals in official statistics: NSIs' point of view, Paper available on EMOS website: http://www.crosportal.eu/sites/default/files//NTTS2013fullPaper_241%20Sorvillo.pdf.

SZYMKOWIAK, M., (2010). ESSnet on Small Area Estimation. Report on the analysis of questionnaires used in WP 2, October 2010.

TORABI, M., SHOKOOHI, F., (2015). Non-parametric generalized linear mixed models in small area estimation. Canadian Journal of Statistics Volume 43, Issue 1, pages 82–96, March 2015.

WALLGREN, A., WALLGREN, B., (2007, 2014). Register-based Statistics. Statistical Methods for Administrative Data. John Wiley & Sons. Ltd.

WALLGREN, A., WALLGREN, B., (2013). Quality Assessment in Systems with Registers and Sample Surveys. http://www.statistics.gov.hk/wsc/IPS078-P2-S.pdf.

ZHANG, L.-C., (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica, Vol. 66, No. 1. p. 41–63.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 631–634

ABOUT THE AUTHORS

Bell William R. is the Senior Mathematical Statistician for Small Area Estimation at the U.S. Census Bureau. He has been involved with research and methodological development for the Census Bureau’s Small Area Income and Poverty Estimates program since 1994. Along with small area estimation, other areas in which he has done research and development include time series modeling and seasonal adjustment, census coverage estimation, and national population projections.

Buelens Bart works as a statistician at the Methodology Department of Statistics Netherlands. He received his MSc in Mathematics from the University of Leuven, Belgium, and his PhD from the University of Tasmania, Australia. His research interests include small area estimation, mixed-mode sample surveys, model-based inference, computational statistics and data science.

Burgard Jan Pablo is an Assistant Professor for 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, estimation of regional morbidity figures, and measurement of biodiversity.

Cheng Yang is a Lead Scientist for the Current Population Survey (CPS), American Time Use Survey (ATUS), and Housing Vacancy Survey (HVS) at the U.S. Census Bureau. He is also an Adjunct Professor of Statistics in the George Washington University. He received a Bachelor's degree in Applied Mathematics from East China Normal University, a Master's degree in Applied Statistics from the American University, and a Doctor's degree in Mathematical Statistics from the University of Maryland at College Park. His research interests include statistical modeling, survey methodology, small area estimation, and labor force statistics. In 2011, he has been initiating the U.S. Census Bureau DSMD Distinguished Lecture Series. The program has attracted 40 distinguished speakers and discussants from different universities in the US and abroad, government, and private survey organizations.

Franco Carolina is a Research Mathematical Statistician at the U.S. Census Bureau's Center for Statistical Research and Methodology. She received her PhD in Applied Mathematics from the University of Maryland in College Park in 2012, with a dissertation focused on asymptotic theory and semiparametric inference. Currently, her primary research interests are survey statistics and small area estimation.

Golata Elzbieta is an Associate Professor at the Department of Statistics, Poznan University of Economics & Business. Her main research domain is demography, labour market and social statistics, especially territorial differentiation in demographic processes and labour market situation. She is also interested in data quality assessment, particularly population census and other demographic estimates. Regional statistics with special emphasis on survey sampling and small area estimation applied to data of economic activity of population is also one of her fields of study. Professor Golata is an active member of many scientific professional bodies.

Ha Neung S. works as a data scientist at Nielsen Company. He received his PhD in Applied Statistics from University of Maryland in College Park in 2013. The title of his PhD thesis is “Hierarchical Bayesian Estimation of Small Area Means Using Complex Survey Data.” He has completed his postdoctoral program at the Statistical and Applied Mathematical Sciences Institute and the National Institute of Statistical Sciences. His main research areas are small area estimation, Bayesian modelling and complex survey inference.

Lahiri Partha is Professor of the Joint Program in Survey Methodology (JPSM) at the University of Maryland at College Park, and an Adjunct Research Professor of the Institute of Social Research, University of Michigan, Ann Arbor. Prior to coming to Maryland, Dr. Lahiri was the Milton Mohr Distinguished Professor of Statistics at the University of Nebraska-Lincoln. His research interests include survey sampling, official statistics, and small-area estimation. Dr. Lahiri has served on a number of advisory committees, including the U.S. Census Advisory committee and U.S. National Academy panel. Over the years Dr. Lahiri advised various local and international organizations such as the United Nations Development Program, World Bank, Gallup Organization. Dr. Lahiri is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics and an elected member of the International Statistical Institute.

Luna Angela is a Lecturer at the Department of Social Statistics and Demography, University of Southampton. Her research interests are small area estimation, survey sampling and official statistics.

Münnich Ralf is a Full Professor and the Head of the Economic and Social Statistics research group at Trier University. He has lead several large-scale research European 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, as well as Monte-Carlo and microsimulation methods.

Piller Kirsten received her MSc in Medical Statistics from the University of Leicester in 2013. In 2014 she joined the Small Area Estimation Unit at the Office for National Statistics where she has been working on model-based estimates of income and poverty.

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 on the editorial board of Survey Methodology journal. He has served on the Advisory Committee for Statistical Methodology of Statistics Canada since 1985. Professional honours he has received include Honorary Doctorates from the University of Waterloo, Canada (2008) and Catholic University of Sacred Heart, Italy (2013), Waksberg Award for Survey Methodology (2005) and 1993 Gold Medail 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.

Van den Brakel Jan A. is a senior statistician at the Methodology Department at Statistics Netherlands and Extraordinary Professor in Survey Methodology at the Department of Quantitative Economics at Maastricht University School of

Business and Economics. His research interests are in the areas of sampling, design and analysis of experiments, inference in mixed-mode surveys, small area estimation, and time series methods.

Whitworth Alison is the Head of the Small Area Estimation Unit at Office for National Statistics, UK. She has a PhD from Southampton University, Department for Social Statistics, focused on the analysis of demographic and health issues from a longitudinal perspective. She has worked in the development and production of population statistics for approximately 15 years, and has a particular interest in methods for combining Census, administrative and survey data for population outputs.

Zhang Li-Chun is a Professor of Social Statistics at S3RI/Department of Social Statistics and Demography, University of Southampton, and Senior Researcher at Statistics Norway. He has worked and published on a number of subjects in Official Statistics. These include sampling design and coordination, sample survey estimation, non-response, measurement errors, small area estimation, index number calculations, editing and imputation, register-based statistics, statistical matching, etc. He has participated in the EU framework projects EURAREA, DACSEIS, RISQ and BLUE-ETS, and the ESSnet projects Small Area Estimation and Data Integration.

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 635–638

ACKNOWLEDGEMENTS TO REVIEWERS

The Editor and Editorial Board of Statistics in Transition new series wish to thank the following persons who served from 31 December 2014 to 31 December 2015 as peer-reviewers of manuscripts for the ***Statistics in Transition new series – Volume 16, Numbers 1–4***. Since the last issue is actually a joint issue of the *Statistics in Transition new series* and ***Survey Methodology***, the Editor of *Survey Methodology* also wishes to thank the people who have provided help or served as referees for one or more papers in this combined issue. The work of the authors in this special joint issue has benefited from their feedback.

Akpanta A. C., Abia State University, Uturu, Nigeria

Breidt F. Jay, Colorado State University, USA

Diallo Mamadou, Westat, USA

Chambers Raymond, University of Wollongong, AU

Chandra Hukum, Indian Institute for Agricultural Statistics Research, India

Christofakis Manolis, Panteion University of Political & Social Sciences, Greece

Dehnel Grażyna, Poznan University of Economics, Poland

Dihidar Kajal, Indian Statistical Institute Kolkata, India

Domański Czesław, University of Lodz, Poland

Gabler Siegfried, GESIS, Mannheim, Germany

Getka-Wilczyńska Elżbieta, Warsaw School of Economics, Poland

Ghosh Malay, Univeristy of Florida, USA

Hanif Muhammad, Lahore University of Management Sciences, Pakistan

Hidiroglou Michael, Statistics Canada, Canada

Jajuga Krzysztof, Wroclaw University of Economics, Poland

Jędrzejczak Alina, University of Lodz, Poland

Kalton Graham, WESTAT, and University of Maryland, USA

Kordos Jan, Warsaw Management Academy, and Central Statistical Office of Poland

Kosiorowski Daniel, Cracow University of Economics, Poland

Krzyśko Mirosław, Adam Mickiewicz University, Poznan, Poland

Lapiņš Jānis, Bank of Latvia, Riga, Latvia

Lehtonen Risto, University of Helsinki, Finland

Liberda Barbara, University of Warsaw, Poland

Longford Nicholas T., Universitat Pompeu Fabra, Barcelona, Spain

Łapczyński Mariusz, Cracow University of Economics, Poland

Mackie Christopher, Committee on National Statistics, National Academies – Washington, DC., USA

Małecka Marta, University of Lodz, Poland

Masik Grzegorz, University of Gdansk, Poland

Młodak Andrzej, Statistical Office Poznan, Poland

Molina Isabel, Universidad Carlos III de Madrid, Spain

Morales Domingo, Universidad Miguel Hernández de Elche, Spain

Mussini Mauro, University of Verona, Italy

Ochocki Andrzej, Cardinal Stefan Wyszynski University in Warsaw, Poland

Okrasa Włodzimierz, Cardinal Stefan Wyszynski University in Warsaw, and Central Statistical Office of Poland

Ramos Raul, University of Barcelona, Spain

Ranalli Maria Giovanna, University of Perugia, Italy

Rao Jon, Carleton University, Canada

Rao Talluri J., Indian Statistical Institute, India

Salvati Nicola, University of Pisa, Italy

Schmid Timo, Berlin Free University, Germany

Seiss Mark, Virginia Tech, USA

Shukla Rakesh, University of Cincinnati, USA

Suchecka Jadwiga, University of Lodz, Poland

Swain A. K. P. C., Utkal University, Bhubaneswar, India

Szymkowiak Marcin, Poznań University of Economics, Poland

Śleszyński Przemysław, Institute of Geography and Spatial Organization PAS, Poland

Tarka Piotr, Poznan University of Economics, Poland

Traat Imbi, University of Tartu, Estonia

Trzpiot Grażyna, University of Economics in Katowice, Poland

Veijanen Ari, Statistics Finland, Finland

Vishwakarma Gajendra K., Indian School of Mines, India

Wiśniewski Jerzy, Nicolaus Copernicus University in Torun, Poland

Wołyński Waldemar, Adam Mickiewicz University of Poznan, Poland

Wywiał Janusz L., University of Economics in Katowice, Poland

Zieliński Wojciech, Warsaw University of Life Sciences, Poland

STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 639–642

INDEX OF AUTHORS, VOLUME 16, 2015

Allin P., *Official Statistics on Personal Well-Being: Some Reflections on the Development and Use of Subjective Well-Being Measures in the UK*

Arcagni A., see under Fattore M.

Bell W. R., see under Franco C.

Bonnéry D., **Cheng Y.**, **Ha N. S.** and **Lahiri P.**, *Triple-goal estimation of unemployment rates for U.S. states using the U.S. Current Population Survey data*

Budnikas G., *Computerised recommendations on e-transaction finalisation by means of machine learning*

Buelens B., see under Van den Brakel J.

Burgard J. P. and **Münnich R.,** *SAE teaching using simulations*

Chambers R. and **Ghosh M.,** *From the Guest Editors (Part 1)*

Cheng Y., see under Bonnéry D.

Clark R. G., see under Molefe W. B.

Das G., see under Ralte Z.

Das M., see under Swain A. K. P. C.

Dehnel G., *Robust regression in monthly business survey*

Dwivedi A. K., Mallawaarachchi I., Figueroa-Casas J. B., Morales A. M. and **Tarwater P.,** *Multinomial logistic regression approach for the evaluation of binary diagnostic test in medical research*

Elkasabi M. A., Heeringa S. G. and **Lepkowski J. M.**, *Joint Calibration Estimator for dual frame surveys*

Fattore M., Maggino F. and **Arcagni A.,** *Exploiting Ordinal Data for Subjective Well-Being Evaluation*

Figueroa-Casas J. B., see under Dwivedi A. K.

Franco C. and **Bell W. R.,** *Borrowing information over time in binomial/logit normal models for small area estimation*

Ghosh M., see under Chambers R.

Gołata E., *SAE education challenges to academics and NSI*

Górecki T., Krzyśko M. and **Wołyński W.,** *Classification problems based on regression models for multi-dimensional functional data*

Górna J., see under Górna K.

Górna K. and **Górna J.,** *Analysis of convergence of European regions with the use of composite index*

Ha N. S., see under Bonnéry D.

Heeringa S. G., see under Elkasabi M. A.

Hidiroglou M. and **Okrasa W.,** *From the Editors*

Homa M. and **Mościbrodzka M.,** *Application of multifactorial market-timing models to assess risk and effectiveness of equity-linked insurance funds in Poland*

Kalton G., Mackie C. and **Okrasa W.,** *From the Editors*

Karlberg F., *Small area estimation for skewed data in the presence of zeroes*

Koźmiński A. K., Noga A., Piotrowska K. and **Zagórski K.,** *Operationalization and Estimation of Balanced Development Index for Poland 1999–2016*

Krzyśko M., see under Górecki T.

Lahiri P., see under Bonnéry D**.**

Lehtonen R. and **Traat I.,** *In Memoriam Gunnar Kulldorff (1927–2015)*

Lepkowski J. M., see under Elkasabi M. A.

Lone H. A. and **Tailor R.,** *Improved separate ratio and product exponential type estimators in the case of post-stratification*

Longford N. T., *Policy-oriented inference and the analyst-client cooperation. An example from small-area statistics*

Luna A., Zhang Li-Chun, Whitworth A. and **Piller K.,** *Small area estimates of the population distribution by ethnic group in England: a proposal using structure preserving estimators*

Mackie C. and **Smith C.,** *Conceptualizing Subjective Well-Being and its Many Dimensions – Implications for Data Collection in Official Statistics and for Policy Relevance;* see also under Kalton G.

Maggino F., see under Fattore M.

Mallawaarachchi I., see under Dwivedi A. K.

Markowska M., *A measure for regional resilience to economic crisis*

Mishra G. C., see under Singh A.

Molefe W. B., Shangodoyin D. K. and **Clark R. G.,** *An approximation to the optimal subsample allocation for small areas*

Morales A. M., see under Dwivedi A. K.

Münnich R., see under Burgard J. P.

Noga A., see under Koźmiński A. K.

Obrębalski M. and **Walesiak M.,** *Functional structure of Polish regions in the period 2004–2013 – measurement via HHI Index, Florence's coefficient of localization and cluster analysis*

Okrasa W., *From the Editor;* see under Kalton G.; see under Hidiroglou M.

Pietrzyk R. and **Rokita P.,** *Stochastic goals in financial planning for a two person household*

Piotrowska K., see under Koźmiński A. K.

Qu X., see under Xing Z.

Piller K., see under Luna A.

Ralte Z. and **Das G.,** *Ratio – to - regression estimator in successive sampling using one auxiliary variable*

Rao J. N. K., *Inferential issues in model-based small area estimation: some new developments*

Rokita P., see under Pietrzyk R.

Shangodoyin D. K., see under Molefe W. B.

Singh A. and **Mishra G. C.,** *Application of Box-Jenkins method and Artificial Neural Network procedure for time series forecasting of prices*

Smith C., see under Mackie C.

Smith D. M., *Using the Day Reconstruction Method to Quantify Time Spent Suffering among Older Adults with Chronic Pain*

Swain A. K. P. C. and **Das M.,** *Some classes of modified ratio type estimators in sample surveys*

Tailor R., see under Lone H. A.

Tarwater P., see under Dwivedi A. K.

Tinkler L., *The Office for National Statistics Experience of Collecting and Measuring Subjective Well-Being*

Traat I., see under Lehtonen R.

Van den Brakel J. and **Buelens B.,** *Covariate selection for small area estimation in repeated sample surveys*

Walesiak M., see under Obrębalski M.

Wilk J., *Using symbolic data in gravity model of population migration to reduce modifiable areal unit problem (MAUP)*

Whitworth A., see under Luna A.

Wołyński W., see under Górecki T.

Xing Z. and **Qu X.,** *An Initial Research on Output Well-Being Index Applied to Residents in Wealthy Counties from China*

Zagórski K., see under Koźmiński A. K.

Zhang Li-Chun, see under Luna A.