Rao, J. N. K.

Article

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

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Rao, J. N. K. (2015) : INFERENTIAL ISSUES IN MODEL-BASED SMALL AREA ESTIMATION: SOME NEW DEVELOPMENTS, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, NY, Vol. 16, Iss. 4, pp. 491-510, https://doi.org/10.21307/stattrans-2015-029

This Version is available at:
https://hdl.handle.net/10419/207787

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

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

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 $\hat{P}_i$ 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 $\overline{Y}_i$. 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

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

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 $t_i^{*} = (\hat{\theta}_{i*}^{EB} - \hat{\theta}_i^{EB}) / [g_{1i}(\hat{\sigma}_{v*}^2)]^{1/2}$. 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 $\hat{\theta}_i^{EB}$ 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

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

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 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.