# New Important Developments in Small Area Estimation

*Danny Pfeffermann*

Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK

Department of Statistics, Hebrew University of Jerusalem, Jerusalem, 91905, Israel

msdanny@soton.ac.il

## Abstract

The problem of small area estimation (SAE) is how to produce reliable estimates of characteristics of interest such as means, counts, quantiles, etc., for areas or domains for which only small samples or no samples are available, and how to assess their precision. The purpose of this paper is to review and discuss some of the new important developments in small area estimation methods. Rao (2003) wrote a very comprehensive book, which covers all the main developments in this topic until that time. A few review papers have been written after 2003 but they are limited in scope. Hence, the focus of this review is on new developments in the last 7-8 years but to make the review more self-contained, I also mention shortly some of the older developments. The review covers both design-based and model-dependent methods, with the latter methods further classified into frequentist and Bayesian methods. The style of the paper is similar to the style of my previous review on SAE published in 2002, explaining the new problems investigated and describing the proposed solutions, but without dwelling on theoretical details, which can be found in the original articles. I hope that this paper will be useful both to researchers who like to learn more on the research carried out in SAE and to practitioners who might be interested in the application of the new methods.

**Key words:** Benchmarking, Calibration, Design-based methods, Empirical likelihood, Informative sampling, Matching priors, Measurement errors, Model checking, M-quantile, Ordered means, Outliers, Poverty mapping, Prediction intervals, Prediction MSE, Spline regression, Two part model.

## ACKNOWLEDGMENTS

I am very grateful to three reviewers for very constructive comments which enhanced the discussion and coverage of this review very significantly.

## 1. PREFACE

The problem of small area estimation (SAE) is how to produce reliable estimates of characteristics of interest such as means, counts, quantiles, etc., for areas or domains for which only small samples or no samples are available. Although the point estimators are usually of first priority, a related problem is how to assess the estimation (prediction) error. The great importance of SAE stems from the fact that many new programs, such as fund allocation for needed areas, new educational or health programs and environmental planning rely heavily on these estimates. SAE techniques are also used in many countries to test and adjust the counts obtained from censuses that use administrative records.

In 2002 I published a review paper with a similar title (Pfeffermann, 2002). At that year small area estimation (SAE) was flourishing both in research and applications, but my own feeling then was that the topic has been more or less exhausted in terms of research and that it will just turn into a routine application in sample survey practice. As the past 9 years show, I was completely wrong and not only that research in this area is accelerating, it now involves some of the best known statisticians, who otherwise are not involved in survey sampling theory or applications. The diversity of new problems investigated is overwhelming, and the solutions proposed are not only elegant and innovative, but also very practical.

Rao (2003) published a comprehensive book on SAE that covers all the main developments in this topic until that time. The book was written about ten years after the review paper of Rao and Ghosh (1994), published in *Statistical Science*, which simulated much of the early research in SAE. Since 2003, a few other review papers have been published; see, e.g., Rao (2005, 2008), Jiang and Lahiri (2006), Datta (2010) and Lehtonen and Veiganen (2009). Notwithstanding, SAE is applied so broadly that I decided that the time is ripe for a new comprehensive review that focuses on the main developments in the last 7-8 years that I am aware of, and which are hardly covered in the review papers mentioned above. The style of the paper is similar to the style of my previous review, explaining the problems investigated and describing the proposed solutions, but without dwelling on theoretical details, which can be found in the original articles. For further clarity and to make the paper more self-contained, I start with a short background and overview some of the 'older' developments. I hope that this paper will be useful to researchers who wish to learn on the research carried out in SAE and to practitioners who might be interested in applying the new methods.

## 2. SOME BACKGROUND

The term ‘SAE’ is somewhat confusing, since it is the size of the sample in the area that causes estimation problems and not the size of the area. Also, the ‘areas’ are not necessarily geographical districts and may define another grouping, such as socio-demographic groups or types of industry, in which case they are often referred to as domains. Closely related concepts in common use are ‘poverty mapping’ or ‘disease mapping’, which amount to SAE of poverty measures or disease incidence and then presenting the results on a map, with different colors defining different levels (categories) of the estimators. What is common to most small area estimation problems is that point estimators and error measures are required for every area separately, and not just as an average over all the areas under consideration.

SAE methods can be divided broadly into ‘design-based’ and ‘model- based’ methods. The latter methods use either the frequentist approach or the full Bayesian methodology, and in some cases combine the two, known in the SAE literature as ‘empirical Bayes’. Design-based methods often use a model for the construction of the estimators (known as ‘model assisted’), but the bias, variance and other properties of the estimators are evaluated under the randomization (design-based) distribution. The randomization distribution is the distribution over all possible samples that could be selected from the target population of interest under the sampling design used to select the sample, with the population measurements considered as fixed values (parameters). Model-based methods on the other hand usually condition on the selected sample and the inference is with respect to the underlying model.

A common feature to design- and model-based SAE is the use of auxiliary covariate information, as obtained from large surveys and/or administrative records such as censuses and registers. Some estimators only require knowledge of the covariates for the sampled units and the true area means of these covariates. Other estimators require knowledge of the covariates for every unit in the population. The use of auxiliary information for SAE is vital because with the small sample sizes often encountered in practice, even the most elaborated model can be of little help if it does not involve a set of covariates that provide ample information on the small area quantities of interest.

## 3. NOTATION

Consider a population $U$ of size $N$, divided into $M$ exclusive and exhaustive areas $U_1 \cup ... \cup U_M$ with $N_i$ units in area $i$, $\sum_{i=1}^{M} N_i = N$. Suppose that samples are available for $m \le M$ of the areas and let $s = s_1 \cup ... \cup s_m$ define the overall sample, where $s_i$ of size $n_i$ is the sample observed for area $i$, $\sum_{i=1}^{m} n_i = n$. Note that $n_i$ is random unless a planned sample of fixed size is taken in that area. Let $y$ define the characteristic of interest and denote by $y_{ij}$ the response value for unit $j$ belonging to area $i$, $i = 1,...,M$; $j = 1...N_i$ with sample means $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$, where we assume without loss of generality that the sample consists of the first $n_i$ units. We denote by $\mathrm{x}_{ij} = (x_{1ij},...,x_{pij})'$ the covariate values associated with unit $(i, j)$ and by $\bar{\mathrm{x}}_i = \sum_{j=1}^{n_i} \mathrm{x}_{ij} / n_i$ the column vector of sample means. The corresponding vector of true area means is $\bar{X}_i = \sum_{j=1}^{N_i} \mathrm{x}_{ij} / N_i$. The area target quantity is denoted by $\theta_i$; for example, $\theta_i = \bar{Y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i$, the response area mean. Estimating a proportion is a special case where $y_{ij}$ is binary. In other applications $\theta_i$ may represent a count or a quantile.

## 4. DESIGN-BASED METHODS

### 4.1 Design-Based Estimators in Common Use

A recent comprehensive review of design-based methods in SAE is provided by Lehtonen and Veijanen (2009). Here I only overview some of the basic ideas. Suppose that the sample is selected by simple random sampling without replacement (SRSWOR) and that the target quantities of interest are the means $\bar{Y}_i$. Estimation of a mean contains as special cases the estimation of a proportion and the estimation of the area distribution $F_i(t) = \sum_{j \in U_i} v_{ij} / N_i$, in which case $v_{ij} = I(y_{ij} \le t)$, where $I(A)$ is the indicator function. Estimators of the percentiles of the area distribution are commonly obtained from the estimated distribution.

If no covariates are available the *direct* design-unbiased estimator of the area mean and its conditional design variance over the *randomization* distribution given $n_i$ are given by,

(4.1) $$\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i \;\; ; \;\; V_D[\bar{y}_i \mid n_i] = (S_i^2 / n_i)[1 - (n_i / N_i)],$$

where $S_i^2 = \sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_i)^2 / (N_i - 1)$. The term 'direct' is used to signify an estimator that only uses the data available for the target area at the specific time of interest. The variance $V_D[\bar{y}_i \mid n_i]$ is $O(1/n_i)$ and for small $n_i$ it is usually large, unless $S_i^2$ is sufficiently small.

Next suppose that covariates $\mathrm{x}_{ij}$ are also observed with $x_{1ij} \equiv 1$. An estimator in common use that utilizes the covariate information is the *synthetic* estimator,

$$\hat{\bar{Y}}_{reg,i}^{syn} = \bar{X}_i' \hat{B} = \frac{1}{N_i} \sum_{j=1}^{N_i} (x_{ij}' \hat{B}), \tag{4.2}$$

where $\hat{B} = [\sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} x_{ij}']^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} y_{ij}$ is the ordinary least square (OLS) estimator. Notice that under SRSWOR, $\hat{B}$ is approximately design-unbiased and consistent for the vector $B$ of regression coefficients computed from all the population values, irrespective of whether a linear relationship between $y$ and x exists in the population. Design-unbiasedness and consistency are with respect to the randomization distribution, letting $N$ and $n$ increase to infinity in a proper way. An estimator is approximately design-unbiased if the randomization bias tends to zero as the sample size increases. The term **"**synthetic" refers to the fact that an (approximately) design-unbiased estimator computed from all the areas ($\hat{B}$ in the present case) is used for every area separately, assuming that the areas are 'homogeneous' with respect to the quantity being estimated. Thus, synthetic estimators borrow information from other 'similar areas' and they are therefore *indirect* estimators.

The obvious advantage of the synthetic estimator over the simple sample mean or any other direct estimator such as the regression estimator $\hat{\bar{Y}}_{reg,i} = \bar{y}_i + (\bar{X}_i - \bar{x}_i)' \hat{B}_i$, where $\hat{B}_i$ is computed only from the data observed for area $i$, is that $Var_D(\hat{\bar{Y}}_{reg,i}^{syn}) = O(1/n)$, and $n = \sum_{i=1}^{m} n_i$ is usually large. The use of the synthetic estimator is motivated ("assisted") by a linear regression model of $y$ on x in the population with a common vector of coefficients. However, for $\mathrm{x}_{1ij} \equiv 1$, $E_D(\hat{\bar{Y}}_{reg,i}^{syn} - \bar{Y}_i) \cong -\bar{X}_i'(B_i - B)$, where $B_i$ is the OLS computed from all the population values in area $i$. Thus, if in fact different regression coefficients $B_i$ operate in different areas, the synthetic estimator may have a large bias. When the sample is selected with unequal probabilities, the OLS estimator $\hat{B}$ in (4.2) is commonly replaced by the probability weighted (PW) estimator $\hat{B}_{pw} = [\sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} x_{ij} x_{ij}']^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} x_{ij} y_{ij}$, where $\{w_{ij} = 1/\Pr[(i,j) \in s]\}$ are the base sampling weights.

In order to deal with the possible large bias of the synthetic estimator, it is common to estimate the bias and then subtract it from the synthetic estimator. The resulting *survey regression* estimator takes the form,

$$\hat{\bar{Y}}_i^{S-R} = \bar{X}_i'\hat{B}_{pw} + \frac{1}{N_i}\sum_{j=1}^{n_i} w_{ij}(y_{ij} - x_{ij}'\hat{B}_{pw}) = \hat{\bar{Y}}_{i,H-T} + (\bar{X}_i - \hat{\bar{X}}_{i,H-T})'\hat{B}_{pw}, \tag{4.3}$$

where $(\hat{\bar{Y}}_{i,H-T}, \hat{\bar{X}}_{i,H-T})$ are the Horvitz-Thompson (H-T) estimators of $(\bar{Y}_i, \bar{X}_i)$. The estimator (4.3) is approximately design-unbiased and performs well when the covariates have good predictive power, but the variance is back to order $O(1/n_i)$. The variance is often reduced by multiplying the bias correction, $\sum_{j=1}^{n_i} w_{ij}(y_{ij} - x_{ij}'\hat{B}_{pw})/N_i$ by $N_i / \sum_{j=1}^{n_i} w_{ij} = N_i / \hat{N}_i$.

A compromise between the possibly large bias of the synthetic estimator and the possibly large variance of the survey regression estimator is achieved by taking a linear combination of the two. The resulting *combined* (*composite*) estimator is defined as,

$$\hat{\bar{Y}}_i^{COM} = \delta_i \hat{\bar{Y}}_i^{S-R} + (1-\delta_i)\hat{\bar{Y}}_{reg,i}^{syn}; \quad 0 \le \delta_i \le 1. \tag{4.4}$$

Ideally, the coefficient $\delta_i$ should be chosen to minimize the mean square error (MSE) of $\hat{\bar{Y}}_i^{COM}$, but assessing accurately the bias of the synthetic estimator for a given area is usually impossible. Hence, it is common to let $\delta_i$ depend on the sample size $n_i$ in the area, such that the larger $n_i$, the larger is $\delta_i$. See Rao (2003) for review of methods of specifying $\delta_i$.

### 4.2 Some New Developments in Design-Based Small Area Estimation

A general class of estimators is obtained by calibrating the base sampling weights $w_{ij}$. Suppose that the population can be partitioned into $C$ calibration groups $U = U_{(1)} \cup ... \cup U_{(C)}$ with known totals $t_{x(c)}$ of the auxiliary variables in the groups, such that each area $U_i$ belongs to one of the groups. Let $s = s_{(1)} \cup ... \cup s_{(C)}$ define the respective partitioning of the sample. In a special case $C = 1$ and $U_{(1)} = U$. The *calibrated* estimator of the mean $\bar{Y}_i$ is computed as,

$$\hat{\bar{Y}}_i^{cal} = \sum_{j=1}^{n_i} w_{ij}^c y_{ij} / N_i \;;\; \sum_{i,j\in s_{(c)}} w_{ij}^c x_{ij} = t_{x(c)}. \tag{4.5}$$

The calibration weights $\{w_{ij}^c\}$ are chosen so that they minimize an appropriate distance from the base weights $\{w_{ij}\}$, subject to satisfying the constraints $\sum_{i,j\in s_{(c)}} w_{ij}^c x_{ij} = t_{x(c)}$. For example, when using the distance $\chi^2 = \sum_{i,j\in s_{(c)}} (w_{ij}^c - w_{ij})^2 / w_{ij}$ and $x_{1ij} \equiv 1$, the calibrated weights are,

(4.6) $$w_{ij}^{c} = w_{ij} g_{ij}\,; \quad g_{ij} = \{1 + (t_{\mathrm{x(c)}} - \hat{t}_{\mathrm{x(c),H\text{-}T}})' [\sum_{i,j \in S_{(c)}} w_{ij} \mathrm{x}_{ij} \mathrm{x}'_{ij}]^{-1} \mathrm{x}_{ij}\},$$

where $\hat{t}_{\mathrm{x(c),H\text{-}T}}$ is the H-T estimator of the total $t_{\mathrm{x}(c)}$. When $U_c = U_i$ (the calibration group is the domain), $\hat{\bar{Y}}_i^{cal}$ is the familiar *generalized regression* (GREG) estimator in the domain.

Calibration of the sampling weights is in broad use in sample survey practice not only for SAE. See Kott (2009) for a recent comprehensive review and discussion. The rationale of the use of calibrated estimators in SAE is that if $y$ is approximately a linear combination of x in the domains belonging to $U_{(c)}$, then $\bar{Y}_i \cong \bar{X}_i' B_{(c)}$ for domains $i \in U_c$, and since $\sum_{i,j \in s_{(c)}} w_{ij}^{c} \mathrm{x}_{ij} = t_{\mathrm{x(c)}}$, $\hat{\bar{Y}}_i^{cal} = \sum_{j=1}^{n_i} w_{ij}^{c} y_{ij} / N_i$ is expected to be a good estimator of $\bar{Y}_i$. Indeed, the advantage of the estimator (4.5) over (4.3) is that it is assisted by a model that only assumes common regression coefficients within the groups $U_{(c)}$, and not for all the domains, as implicitly assumed by the estimator (4.3). The estimator (4.5) is approximately design-unbiased irrespective of any model, but $Var_D(\hat{\bar{Y}}_i^{cal} \mid n_i) = O(1/n_i)$, which may still be large.

Another way of calibrating the weights is by use of *instrumental variables* (Estevao and Särndal, 2004, 2006). Denote the vector of instrument values for unit $(i, j)$ by $h_{ij}$. The calibrated weights are defined as,

(4.7) $$w_{ij}^{ins} = w_{ij}(1 + g_c' h_{ij})\ ;\ g_c' = (t_{\mathrm{x(c)}} - \hat{t}_{\mathrm{x(c),H\text{-}T}})' [\sum_{i,j \in S_{(c)}} w_{ij} h_{ij} \mathrm{x}'_{ij}]^{-1}\,.$$

Note that the instrument values need only be known for the sampled units in $s_{(c)}$ and that $\sum_{i,j \in s_{(c)}} w_{ij}^{ins} \mathrm{x}_{ij} = t_{c\mathrm{x}}$, thus satisfying the same constraints as before. The calibrated estimator of $\bar{Y}_i$ is now $\hat{\bar{Y}}_{i,ins}^{cal} = \sum_{j=1}^{n_i} w_{ij}^{ins} y_{ij} / N_i$. When h=x, $w_{ij}^{ins} = w_{ij}^{c}$. The use of instruments replaces the search for an appropriate distance function by imposing a structure on the calibration weights, and it allows in principle finding the best weights in terms of minimizing an approximation to the variance of the calibrated estimator. However, as noted by Estevao and Särndal (2006), the optimal weights depend on unknown population quantities which, when estimated from the sample, may yield unstable estimators. See Kott (2009) for further discussion.

The synthetic estimator (4.2), the survey regression estimator (4.3) and the various calibrated estimators considered above are all assisted by models that assume a linear relationship between y and x. These estimators only require knowledge of the covariates for the sampled units, and the area (or group) totals of these covariates. Lehtonen *et al*. (2003, 2005) consider the use of generalized linear models (GLM), or even generalized linear mixed

models (GLMM) as the assisting models, which require knowledge of the covariates for every element in the population. Suppose that $E_M(y_{ij}) = f(\mathrm{x}_{ij};\psi)$ for some nonlinear function $f(\cdot)$ with an unknown vector parameter $\psi$, where $E_M(\cdot)$ defines the expectation under the model. A simple important example is where $f(\mathrm{x}_{ij};\psi)$ is the logistic function. Estimating $\psi$ by the *pseudo-likelihood* (PL) approach yields the estimator $\hat{\psi}_{pl}$ and predicted values $\{\hat{y}_{ij} = f(\mathrm{x}_{ij};\hat{\psi}_{pl})\}$. The PL approach consists of estimating the likelihood equations that would be obtained in case of a census by the corresponding H-T estimators (or weighting each score function by its sampling weight), and then maximizing the resulting estimated equations. The synthetic and "generalized GREG" estimators are computed as,

$$(4.8) \qquad \hat{\bar{Y}}^{syn}_{GLM,i} = \frac{1}{N_i}\sum\nolimits_{j=1}^{N_i} f(\mathrm{x}_{ij};\hat{\psi}_{pl});\ \ \hat{\bar{Y}}^{GREG}_{GLM,i} = \hat{\bar{Y}}^{syn}_{GLM,i} + \frac{1}{N_i}\sum\nolimits_{j=1}^{n_i} w_{ij}[y_{ij} - f(\mathrm{x}_{ij};\hat{\psi}_{pl})].$$

A further extension is to include random area effects in the assisting model assuming, $E_M(y_{ij} \mid \mathrm{x}_{ij}, u_i) = f(\mathrm{x}_{ij}, u_i;\psi^*)$, $E_M(u_i) = 0$, $Var_M(u_i) = \sigma_u^2$. Estimation of the fixed parameters $\psi^*, \sigma_u^2$ and the random effects $u_i$ is now under the model, ignoring the sampling weights. The extended synthetic and generalized GREG estimators are defined similarly to (4.8), but with $f(\mathrm{x}_{ij};\hat{\psi}_{pl})$ replaced by $f(\mathrm{x}_{ij},\hat{u}_i;\hat{\psi}^*)$. For sufficiently large sample size $n_i$, the extended generalized GREG is approximately design-unbiased for the true area mean but it is not clear how to estimate the design (randomization) variance in this case in a way that accounts for the prediction of the random effects. Torabi and Rao (2008) compare the MSE of model-based predictors and a GREG assisted by a linear mixed model (LMM).

Jiang and Lahiri (2006a) propose the use of model-dependent estimators that are design-consistent under the randomization distribution as the area sample sizes increase. The basic idea is to model the direct estimators $\hat{\bar{Y}}_{iw} = \sum_{j=1}^{n_i} w_{ij} y_{ij} / \sum_{j=1}^{n_i} w_{ij}$ instead of the individual observations $y_{ij}$, and then employ the empirical best predictor of the area mean under the model. The authors consider the general two-level model $E_M[\hat{\bar{Y}}_{iw} \mid u_i] = \xi_i = \xi(u_i, \hat{\bar{X}}_{iw};\psi)$, where the $u_i$'s are independent random area effects with zero mean and variance $\sigma_u^2$, $\hat{\bar{X}}_{iw} = \sum_{j=1}^{n_i} w_{ij}\mathrm{x}_{ij} / \sum_{j=1}^{n_i} w_{ij}$ and $\xi(\cdot)$ is some known function with unknown parameters $\psi$. The empirical best predictor is the best predictor under the model (minimum expected quadratic loss), but with the parameters $\psi$ replaced by model consistent estimators;

$\hat{\bar{Y}}_i^{EBP} = E_M(\xi_i \mid \hat{\bar{Y}}_{iw}; \hat{\psi})$. The estimator is shown to be model-consistent under correct model specification and design-consistent for large $n_i$ even if the model is misspecified, thus robustifying the estimation. The authors develop estimators of the prediction mean squared error (PMSE) for bounded sample sizes, with bias of desired order $o(1/m)$, where $m$ is the number of sampled areas. The PMSE is computed with respect to the model holding for the individual observations and over the randomization distribution. The use of design consistent estimators in SAE is somewhat questionable because of the small sample sizes in some or all of the areas, but it is nonetheless a desirable property. This is so because it is often the case that in some of the areas the samples are large and it is essential that an estimator should work well at least in these areas, even if the model fails. Estimators with large randomization bias even for large samples do not appeal to practitioners.

Chandra and Chambers (2009) propose the use of model-based direct estimators (MBDE). The idea is to fit a model for the population values, compute the weights defining the Empirical Best Linear Unbiased Predictor (EBLUP) of the population total under the model, and then use the weights associated with a given area to compute an almost direct estimator. The model fitted for the population values $Y_U$ is the general linear model,

$$Y_U = X_U\beta + \varepsilon_U;\ E(\varepsilon_U) = 0,\ E(\varepsilon_U\varepsilon_U') = \Sigma = \begin{bmatrix} \Sigma_{ss} & \Sigma_{sr} \\ \Sigma_{rs} & \Sigma_{rr} \end{bmatrix}, \tag{4.9}$$

where $s$ signifies the sample of size $n$ and $r$ signifies the sample-complement of size $(N-n)$. As seen later, the models in common use for SAE defined by (5.1) and (5.3) below are special cases of (4.9). Let $y_s$ denote the column vector of sample outcomes. For known $\Sigma$, the BLUP of the population total $t_y = \sum_{k=1}^{N} y_k$ under the model is,

$$\hat{t}_y^{BLUP} = 1_n' y_s + 1_{N-n}'[X_r\hat{\beta}_{GLS} + \Sigma_{rs}\Sigma_{ss}^{-1}(y_s - X_s\hat{\beta}_{GLS})] = \sum_{k\in s} w_k^{BLUP} y_k\,, \tag{4.10}$$

where $\hat{\beta}_{GLS}$ is the generalized least square estimator. The EBLUP is $\hat{t}_y^{EBLUP} = \sum_{k\in s} w_k^{EBLUP} y_k$, where the EBLUP weights are the same as in (4.10) but with estimated parameters. The MBDE of the true mean in area $i$ is,

$$\hat{\bar{Y}}_i^{MBD} = \sum_{j\in s_i} w_j^{EBLUP} y_i \Big/ \sum_{j\in s_i} w_j^{EBLUP}\,. \tag{4.11}$$

The authors derive estimators for the bias and variance of the MBDE and illustrate its robustness to certain model misspecifications. Note, however, that $\hat{\bar{Y}}_i^{MBD}$ is a ratio estimator and therefore may have a non-negligible bias in areas $i$ with small sample size.

All the estimators considered so far assume a given sampling design with random area sample sizes. When the target areas are known in advance, considerable gains in efficiency can be achieved by modifying the sampling design and in particular, by controlling the sample sizes within these areas. In a recent article, Falrosi and Righi (2008) propose a general strategy for multivariate multi-domain estimation that guarantees that the sampling errors of the domain estimators are lower than pre-specified thresholds. The strategy combines the use of a balanced sampling technique and GREG estimation, but extensions to the use of synthetic estimators and model-based estimation are also considered. A successful application of this strategy requires good predictions of weighted sums of residuals featuring in the variance expressions, and it may happen that the resulting overall sample size is far too large, but this is a promising approach that should be studied further.

### 4.3 Pros and Cons of Design-Based Small Area Estimation

The apparent advantage of design-based methods is that the estimation is less dependent on an assumed model, although models are used (assisted) for the construction of the estimators. The estimators are approximately unbiased and consistent under the randomization distribution for large sample sizes within the areas, which as discussed before is a desirable property that protects against possible model misspecification at least in large areas.

Against this advantage stand many disadvantages. Direct estimators generally have large variance due to small sample sizes. The survey regression estimator is approximately unbiased but may likewise be too variable. Synthetic estimators have small variance but are generally biased. Composite estimators have smaller bias than synthetic estimators but larger variance, and it is not obvious how to best choose the weights attached to the synthetic estimator and the GREG (or another direct estimator). Computation of randomization-based confidence intervals generally requires large sample normality assumptions, but the sample sizes in at least some of the areas may be too small to justify asymptotic normality.

Another limitation of design-based inference (not restricted to SAE) is that it does not lend itself to conditional inference, for example, conditioning on the sampled values of the covariates or the sampled clusters in a two-stage sampling design. This again inflates the variance of the estimators. Conditional inference is in the heart of classical statistical inference under both the frequentist and the Bayesian approaches. Last, but not least important limitation of design-based SAE is that there is no founded theory for estimation in areas with no samples. The use of the randomization distribution does not extend to prediction problems, such as the prediction of small area means for areas with no samples. It

is often the case that samples are available for only a minority of the areas but estimators and MSE estimators are required for each of the areas, whether sampled or not.

## 5. MODEL-BASED METHODS

### 5.1 General Formulation

Model-based methods assume a model for the sample data and use the optimal or approximately optimal predictor of the area characteristic of interest under the model. The MSE of the prediction error is likewise defined and estimated with respect to the model. Note that I now use the term 'prediction' rather than estimation because the target characteristics are generally random under the model. The use of models overcomes the problems underlying the use of design-based methods but it is important to emphasize again that even the most elaborated model cannot produce sufficiently accurate predictors when the area sample size is too small and no covariates with good predictive power are available. The use of models raises the question of the robustness of the inference to possible model misspecification, and Sections 6.3-6.6 review studies that deal with this problem from different perspectives. Section 8 considers model selection and diagnostic checking.

Denote by $\theta_i$ the target quantity in area $i$ (mean, proportion,…). Let $y_i$ define the observed responses for area $i$ and $\mathrm{x}_i$ define the corresponding values of the covariates (when available). As becomes evident below, $y_i$ is either a scalar, in which case $\mathrm{x}_i$ is a vector, or $y_i$ is a vector, in which case $\mathrm{x}_i$ is usually a matrix. A typical small area model consists of two parts: The first part models the distribution (or just the moments) of $y_i \mid \theta_i; \psi_{(1)}$. The second part models the distribution (moments) of $\theta_i \mid \mathrm{x}_i; \psi_{(2)}$, linking the $\theta_i$'s to known covariates and to each other. This is achieved by including in the model random effects that account for the variability of the $\theta_i$'s not explained by the covariates. The hyper-parameters $\psi = (\psi_{(1)}, \psi_{(2)})$ are typically unknown and are estimated either under the frequentist approach or under the Bayesian approach with appropriate prior distributions. In some applications the index $i$ may define time, in which case the model for $\theta_i \mid \mathrm{x}_i; \psi_2$ is a time series model.

### 5.2 Models in Common Use

In this section I review briefly three models in common use, as most of the recent developments in SAE relate to these models or extensions of them. For more details see Rao (2003), Jiang and Lahiri (2006), Datta (2009) and the references therein. I assume that the model holding for the sample data is the same as the model holding in the population, so that

there is no sample selection bias. The case of informative sampling of areas or within the selected areas, whereby the sample selection or response probabilities are related to the response variable even after conditioning on the model covariates is considered in Section 7. Notice that in this case the sample model differs from the population model.

5.2.1 *Area level model*

This model is in broad use when the covariate information is only at the area level, so that $\mathrm{x}_i$ is a vector of known area characteristics. The model, studied originally for SAE by Fay and Herriot (1979) is defined as,

(5.1) $$\tilde{y}_i = \theta_i + e_i \;\; ; \;\; \theta_i = \mathrm{x}_i'\beta + u_i ,$$

where $\tilde{y}_i$ denotes the direct sample estimator of $\theta_i$ (for example, the sample mean $\bar{y}_i$ when the sample is selected by SRS) and $e_i$ represents the sampling error, assumed to have zero mean and known design (randomization) variance, $Var_D(e_i) = \sigma_{Di}^2$. The random effects $u_i$ are assumed to be independent with zero mean and variance $\sigma_u^2$. For known $\sigma_u^2$, the best linear unbiased predictor (BLUP) of $\theta_i$ under this model is,

(5.2) $$\hat{\theta}_i = \gamma_i \tilde{y}_i + (1-\gamma_i)\mathrm{x}_i'\hat{\beta}_{GLS} = \mathrm{x}_i'\hat{\beta}_{GLS} + \gamma_i(\tilde{y}_i - \mathrm{x}_i'\hat{\beta}_{GLS}) = \mathrm{x}_i'\hat{\beta}_{GLS} + \hat{u}_i .$$

The BLUP $\hat{\theta}_i$ is in the form of a composite estimate (Eq. 4.4), but with a tuning (shrinkage) coefficient $\gamma_i = \sigma_u^2 / (\sigma_u^2 + \sigma_{Di}^2)$, which is a function of the ratio $\sigma_u^2 / \sigma_{Di}^2$ of the variances of the prediction errors of $\mathrm{x}_i'\beta$ and $\tilde{y}_i$ respectively. The coefficient $\gamma_i$ defines optimally the weight assigned to the synthetic estimator $\mathrm{x}_i'\hat{\beta}_{GLS}$ and $\tilde{y}_i$, unlike the case of design-based estimators where the weight is assigned in a more ad hoc manner. See the discussion below (4.4). Note that the BLUP property does not require specifying the distribution of the error terms beyond the first two moments, and $\hat{\theta}_i$ is also the linear Bayes predictor in this case. Under normality of the error terms and a diffuse uniform prior for $\beta$, $\hat{\theta}_i$ is the Bayesian predictor (posterior mean) of $\theta_i$. For a nonsampled area $k$, the BLUP is now obtained optimally as $\mathrm{x}_k'\hat{\beta}_{GLS}$.

In practice, the variance $\sigma_u^2$ is seldom known and is replaced in $\gamma_i$ and $\hat{\beta}_{GLS}$ by a sample estimate, yielding what is known as the empirical BLUP (EBLUP) under the frequentist approach, or the empirical Bayes (EB) predictor when assuming normality. The latter predictor is the posterior mean of $\theta_i$ but with $\sigma_u^2$ replaced by a sample estimate obtained from the marginal distribution of the direct estimators given the variance. Alternatively, one

may compute the Hierarchical Bayes (HB) predictor by assuming a prior distribution for $\sigma_u^2$ and computing the posterior distribution of $\theta_i$ given the available data, which can be used for computation of the point predictor and a credibility (confidence) interval.

Remark 1. The synthetic estimator $\mathrm{x}_i'\hat{\beta}_{GLS}$, and hence the BLUP $\hat{\theta}_i$ are unbiased predictors under the joint distribution of $y_i$ and $\theta_i$ in the sense that $E(\hat{\theta}_i - \theta_i) = 0$, but are biased when conditioning on $u_i$. The predictor $\hat{\theta}_i$ is also biased under the randomization distribution. Conditioning on $u_i$ amounts to assuming different fixed intercepts in different areas and the unbiasedness of $\hat{\theta}_i$ under the model is achieved by viewing the intercepts as random.

Remark 2. It is often the case that the *linking model* is defined for a transformation of $\theta_i$. For example, Fay and Herriot (1979) actually assume $\log(\theta_i) = \mathrm{x}_i'\beta + u_i$ in (5.1) and use the direct estimator $\tilde{y}_i = \log(\bar{y}_i)$, and then predict $\theta_i$ as $\exp(\tilde{\theta}_i)$, where $\tilde{\theta}_i$ is the BLUP (EBLUP) of $\log(\theta_i)$ under the model. However, $\exp(\tilde{\theta}_i)$ is not the BLUP of $\theta_i = \exp[\log(\theta_i)]$. On the other hand, the EB and HB approaches produce optimal predictors of $\theta_i$ even if the linking model uses a transformation of $\theta_i$, with or without the use of a similar transformation for the direct estimator. In this respect the latter two approaches are more flexible and with wider applicability, but at the expense of requiring further parametric assumptions.

5.2.2 *Nested error unit level model*

This model uses individual observations $y_{ij}$ such that $y_i$ is now a vector and $\mathrm{x}_i$ is generally a matrix. The use of this model for SAE requires that the area means, $\bar{\mathrm{X}}_i = \sum_{j=1}^{N_i} \mathrm{x}_{ij} / N_i$ are known. The model, first proposed for SAE by Battese *et al.* (1988) has the form,

$$y_{ij} = \mathrm{x}_{ij}'\beta + u_i + \varepsilon_{ij}, \tag{5.3}$$

where the $u_i$'s ( random effects) and the $\varepsilon_{ij}$ s (residual terms) are mutually independent with zero means and variances $\sigma_u^2$ and $\sigma_\varepsilon^2$ respectively. Under the model, the true small area means are $\bar{Y}_i = \bar{X}_i'\beta + u_i + \bar{\varepsilon}_i$, but since $\bar{\varepsilon}_i = \sum_{j=1}^{N_i} \varepsilon_{ij} / N_i \cong 0$ for large $N_i$, the target means are often defined as $\theta_i = \bar{X}_i'\beta + u_i = E(\bar{Y}_i \mid u_i)$. For known variances $(\sigma_u^2, \sigma_\varepsilon^2)$, the BLUP of $\theta_i$ is,

$$\hat{\theta}_i = \gamma_i[\bar{y}_i + (\bar{X}_i - \bar{\mathrm{x}}_i)'\hat{\beta}_{GLS}] + (1-\gamma_i)\bar{X}_i'\hat{\beta}_{GLS}, \tag{5.4}$$

where $\hat{\beta}_{GLS}$ is the GLS of $\beta$ computed from all the observations, $\bar{\mathrm{x}}_i = \sum_{j=1}^{n_i} \mathrm{x}_{ij} / n_i$ and $\gamma_i = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2 / n_i)$. For area $k$ with no sample (but known $\bar{X}_k$), the BLUP is $\hat{\theta}_k = \bar{\mathrm{X}}_k' \hat{\beta}_{GLS}$. See Rao (2003) for the BLUP of the means $\bar{Y}_i$ in sampled areas.

The BLUP (5.4) is also the Bayesian predictor (posterior mean) under normality of the error terms and a diffuse uniform prior for $\beta$. Replacing the variances $\sigma_u^2$ and $\sigma_\varepsilon^2$ in $\gamma_i$ and $\hat{\beta}_{GLS}$ by sample estimates yields the corresponding EBLUP or EB predictors. Hierarchical Bayes (HB) predictors are obtained by specifying prior distributions for $\beta$ and the two variances and computing the posterior distribution of $\theta_i$ (or $\bar{Y}_i$) given all the sample observations in all the areas. Remark 1 applies to the BLUP (EBLUP) under this model as well.

5.2.3 *Mixed logistic model*

The previous two models assume continuous responses. Suppose now that $y_{ij}$ is binary taking the values 1 and 0, in which case the small area quantities of interest are usually proportions or counts (say, the proportion or total of unemployed persons in the area). The following generalized linear mixed model (GLMM) considered originally by MacGibbon and Tomberlin (1989) for SAE is in broad use for this kind of problems:

(5.5) $\quad \Pr(y_{ij} = 1 \mid p_{ij}) = p_{ij}\ ;\ logit(p_{ij}) = \mathrm{x}_{ij}'\beta + u_i\ ;\ u_i \sim N(0, \sigma_u^2)$.

The responses $y_{ij}$ are assumed to be conditionally independent given the random effects $u_i$, and likewise for the random effects. The purpose is to predict the true area proportions $p_i = \sum_{j=1}^{N_i} y_{ij} / N_i$. Let $\psi = (\beta, \sigma_u^2)$ denote the model parameters. For this model there is no explicit expression for the best predictor (BP) under a quadratic loss function, that is, for $\hat{p}_i^{BP} = E(p_i \mid y_i, \mathrm{x}_i; \psi)$, but as shown in Jiang and Lahiri (2006b), the BP can be computed (approximated) numerically as the ratio of two one-dimensional integrals. The authors review methods of estimating $\psi$, yielding the empirical BP (EBP) $\hat{p}_i^{EBP} = E(p_i \mid y_i, \mathrm{x}_i; \hat{\psi})$, which is also the EB predictor under the same assumptions. Application of the full HB approach under this model consists of the following basic steps:

1. Specify prior distributions for $\sigma_u^2$ and $\beta$;

2. Compute the posterior distributions of $\beta$, $\sigma_u^2$ and $u_1,...,u_m$ by say, MCMC simulations and draw a large number of realizations $(\hat{\beta}^{(r)}, \sigma_u^{2(r)}, \{\hat{u}_i^{(r)}\})$, $r = 1,...,R$, $i = 1,...,m$, and hence realizations, $y_{ik}^{(r)} \sim p_{ik}^{(r)} = \dfrac{\exp(x_{ik}'\beta^{(r)} + u_i^{(r)})}{1+\exp(x_{ik}'\beta^{(r)} + u_i^{(r)})}$ for $k \notin s_i$;

3. Predict: $\hat{p}_i = (\sum_{j\in s_i} y_{ij} + \sum_{k\notin s_i} \hat{y}_{ik}) / N_i$; $\hat{y}_{ik} = \sum_{r=1}^{R} y_{ik}^{(r)} / R$, $k \notin s_i$.

Writing $\hat{p}_i = \dfrac{1}{R}\sum_{r=1}^{R}(\sum_{j\in s_i} y_{ij} + \sum_{k\notin s_i} \hat{y}_{ik}^{(r)}) / N_i = \dfrac{1}{R}\sum_{r=1}^{R} \hat{p}_i^{(r)}$, the posterior variance is approximated as $\hat{V}_{post}(\hat{p}_i) = \dfrac{1}{R(R-1)}\sum_{r=1}^{R}(\hat{p}_i^{(r)} - \hat{p}_i)^2$.

Ghosh *et al*. (1998) discuss the use of HB SAE for GLMM, covering binary, count, multi-category and spatial data. In particular, sufficient conditions are developed for the joint posterior distribution of the parameters of interest to be proper.

## 6. NEW DEVELOPMENTS IN MODEL-BASED SAE

### 6.1 Estimation of Prediction MSE

As stated in the introduction, an important aspect of SAE is the assessment of the accuracy of the predictors. This problem is solved 'automatically' under the Bayesian paradigm, which produces realizations of the posterior distribution of the target quantities. However, estimation of the prediction MSE (PMSE) and the computation of confidence intervals (C.I.) under the frequentist approach is complicated because of the added variability induced by the estimation of the model hyper-parameters. Prasad and Rao (1990) developed PMSE estimators with bias of order $o(1/m)$, ($m$ is the number of sampled areas), under the linear mixed models (5.1) and (5.2) for the case where the random errors have a normal distribution and the model variances are estimated by the ANOVA method of moments. Datta and Lahiri (2000) extend the estimation of Prasad and Rao to the more general mixed linear model,

(6.1) $$y_i = X_i\beta + Z_i u_i + e_i, \quad i = 1...m,$$

where $X_i$ and $Z_i$ are fixed matrices of order $n_i \times k$ and $n_i \times d$ respectively, and $u_i$ and $e_i$ are independent normally distributed random effects and residual terms of orders $d \times 1$ and $n_i \times 1$ respectively; $u_i \sim N_d(0, Q_i)$, $e_i \sim N_{n_i}(0, R_i)$. The variance matrices are known functions of variance components $\tilde{\psi} = (\tilde{\psi}_1,...,\tilde{\psi}_q)$. The authors develop PMSE estimators with bias of order $o(1/m)$ for the EBLUP obtained when estimating $\beta$ and $\tilde{\psi}$ by MLE or REML. Das *et al*. (2004) extend the model of Datta and Lahiri (2000) by relaxing the assumption of

independence of the error terms between the areas and likewise develop an estimator for the PMSE of the EBLUP when estimating the unknown model parameters by MLE or REML, with bias of order $o(1/m)$. This article contains rigorous proofs. Datta *et al*. (2005) show that for the area level model (5.1), if $\sigma_u^2$ is estimated by the method proposed by Fay and Herriot (1979), it is required to add an extra term to the PMSE estimator to achieve the desired order of bias of $o(1/m)$. See Datta (2009) for a more extensive review of methods of estimating the PMSE of the EBLUP and EB under linear mixed models (LMM).

Estimation of the PMSE under the GLMM is more involved and in what follows I review resampling procedures that can be used in such cases. For convenience, I consider the mixed logistic model (5.5) but the procedures are applicable to other models belonging to this class. The first procedure, proposed by Jiang *et. al* (2002) uses the Jackknife method. Let $\lambda_i = E(\hat{p}_i^{EBP} - p_i)^2$ denote the PMSE, where $p_i = \sum_{j=1}^{N_i} y_{ij} / N_i$ is the true proportion and $\hat{p}_i^{EBP} = E(p_i \mid y_i, \mathrm{x}_i; \hat{\psi})$ is the EBP. The following decomposition holds,

$$\lambda_i = E(\hat{p}_i^{(BP)} - p_i)^2 + E(\hat{p}_i^{(EBP)} - \hat{p}_i^{(BP)})^2 = M_{1i} + M_{2i}, \tag{6.2}$$

where $M_{1i}$ is the PMSE of the BP (assumes known parameter values) and $M_{2i}$ is the contribution to the PMSE from estimating the model parameters, $\psi$. Denote by $\hat{\lambda}_i^{BP}(\hat{\psi})$ the '*naive*' estimator of $M_{1i}$, obtained by setting $\psi = \hat{\psi}$. Let $\hat{\lambda}_i^{BP}(\hat{\psi}_{-l})$ denote the naive estimator when estimating $\psi$ from all the areas except for area $l$, and $\hat{p}_i^{EBP}(\hat{\psi}_{-l})$ denote the corresponding EBP. The Jackknife estimator of PMSE is:

$$\hat{\lambda}_i^{JK} = \hat{M}_{1i} + \hat{M}_{2i}; \quad \begin{matrix} \hat{M}_{1i} = \hat{\lambda}_i^{BP}(\hat{\psi}) - \dfrac{m-1}{m}\sum_{l=1}^{m}[\hat{\lambda}_i^{BP}(\hat{\psi}_{-l}) - \hat{\lambda}_i^{BP}(\hat{\psi})] \\ \hat{M}_{2i} = \dfrac{m-1}{m}\sum_{l=1}^{m}[\hat{p}_i^{EBP}(\hat{\psi}_{-l}) - \hat{p}_i^{EBP}(\hat{\psi})]^2 \end{matrix}. \tag{6.3}$$

Under some regularity conditions, $E(\hat{\lambda}_i^{JK}) - \lambda_i = o(1/m)$, as desired.

The jackknife estimator estimates the unconditional PMSE over the joint distribution of the random effects and the responses. Lohr and Rao (2009) proposed a modification of the jackknife, which is simpler and estimates the conditional PMSE, $E[(\hat{p}_i^{(EBP)} - p_i)^2 \mid y_i]$. Denoting $q_i(\psi, y_i) = Var(\theta_i \mid y_i; \psi)$, the modification consists of replacing $\hat{M}_{1i}$ in (6.3) by $\hat{M}_{1i,c} = q_i(\hat{\psi}, y_i) - \sum_{l \neq i}^{m}[q_i(\hat{\psi}_{-l}, y_i) - q_i(\hat{\psi}, y_i)]$. The modified estimator $\hat{\lambda}_{i,c}^{JK} = \hat{M}_{1i,c} + \hat{M}_{2i}$ has

bias of order $o_p(1/m)$ in estimating the conditional PMSE and bias of order $o(1/m)$ in estimating the unconditional PMSE.

Hall and Maiti (2006) propose estimating the PMSE based on double-bootstrap. For the model (5.5) the procedure consists of the following steps:

1. Generate a new population from the model (5.5) with parameters $\hat{\psi}$ and compute the 'true' area proportions for this population. Compute the EBPs based on new sample data and newly estimated parameters. The new population and sample use the same covariates as the original population and sample. Repeat the process independently $B_1$ times, with $B_1$ sufficiently large. Denote by $p_{i,b_1}(\hat{\psi})$ and $\hat{p}_{i,b_1}^{(EBP)}(\hat{\psi}_{b_1})$ the the 'true' proportions and corresponding EBPs for population and sample $b_1$, $b_1 = 1,...,B_1$. Compute the first-step bootstrap PMSE estimator,

$$\hat{\lambda}_{i,1}^{BS} = \frac{1}{B_1}\sum_{b_1=1}^{B_1}[\hat{p}_{i,b_1}^{(EBP)}(\hat{\psi}_{b_1}) - p_{i,b_1}(\hat{\psi})]^2 . \tag{6.4}$$

2. For each sample drawn in Step 1, repeat the computations of Step 1 $B_2$ times with $B_2$ sufficiently large, yielding new 'true' proportions $p_{i,b_2}(\hat{\psi}_{b_1})$ and EBPs $\hat{p}_{i,b_2}^{(EBP)}(\hat{\psi}_{b_2})$, $b_2 = 1,...,B_2$. Compute the second-step bootstrap PMSE estimator,

$$\hat{\lambda}_{i,2}^{BS} = \frac{1}{B_1}\sum_{b_1}^{B_1}\frac{1}{B_2}\sum_{b_2=1}^{B_2}[\hat{p}_{i,b_2}^{(EBP)}(\hat{\psi}_{b_2}) - p_{i,b_2}(\hat{\psi}_{b_1})]^2 . \tag{6.5}$$

The double-bootstrap PMSE estimator is obtained by computing one of the classical bias corrected estimators. For example,

$$\hat{\lambda}_i^{D-BS} = \begin{cases} \hat{\lambda}_{i,1}^{BS} + (\hat{\lambda}_{i,1}^{BS} - \hat{\lambda}_{i,2}^{BS}), & \text{if } \hat{\lambda}_{i,1}^{BS} \geq \hat{\lambda}_{i,2}^{BS} \\ \hat{\lambda}_{i,1}^{BS}\exp[(\hat{\lambda}_{i,1}^{BS} - \hat{\lambda}_{i,2}^{BS})/\hat{\lambda}_{i,2}^{BS}], & \text{if } \hat{\lambda}_{i,1}^{BS} < \hat{\lambda}_{i,2}^{BS} \end{cases} . \tag{6.6}$$

Notice that whereas the first-step bootstrap estimator (6.4) has bias of order $O(1/m)$, the double-bootstrap estimator has bias of order $o(1/m)$ under some regularity conditions.

Pfeffermann and Correa (2012) develop a general method of bias correction, which models the error of a target estimator as a function of the corresponding bootstrap estimator, and the original estimators and bootstrap estimators of the parameters governing the model fitted to the sample data. This is achieved by drawing at random a large number of plausible parameters governing the model, generating a pseudo original sample for each parameter and bootstrap samples for each pseudo sample, and then searching by a cross validation procedure the best functional relationship among a set of eligible bias correction functions that includes the classical bootstrap bias corrections. The use of this method produces estimators with bias

of correct order and under certain conditions it also permits estimating the MSE of the bias corrected estimator. Application of the method for estimating the PMSE under the model (5.5) in an extensive simulation study outperforms the double-bootstrap and jackknife procedures, with good performance in estimating the MSE of the PMSE estimators.

Remark 3. All the resampling methods considered above are in fact fully parametric since they require computing repeatedly the empirical best predictors under the model.

Chambers *et al*. (2011) develop conditional bias-robust PMSE estimators for the case where the small area estimators can be expressed as weighted sums of sample values. The authors assume that that for unit $j \in U_i$, $y_j = \mathrm{x}'_j\beta_i + e_j$; $E(e_j) = 0$, $Var(e_j) = \sigma_j^2$, $j = 1,...,n_i$ with $\beta_i$ taken as a fixed vector of coefficients, and consider linear estimators of the form $\hat{\theta}_i = \sum_{k\in s} w_{ik} y_k$ with fixed weights $w_{ik}$. Thus, if $\theta_i$ defines the true area mean,

$$
\begin{aligned}
Bias_i &= E(\hat{\theta}_i - \theta_i) = (\sum_{h=1}^{m} \sum_{j\in s_h} w_{ij}\mathrm{x}'_j\beta_h) - \bar{X}_i\beta_i \\
Var_i &= Var(\hat{\theta}_i - \theta_i) = N_i^{-2}(\sum_{h=1}^{m} \sum_{j\in s_h} a_{ij}^2\sigma_j^2 + \sum_{j\in r_i} \sigma_j^2),
\end{aligned}
\tag{6.7}
$$

where $r_i = U_i - s_i$ and $a_{ij} = N_i w_{ij} - I(j \in U_i)$, with $I(\cdot)$ defining the indicator function. Assuming that for $j \in U_i$ $\mu_j = E(y_j \mid \mathrm{x}_j,) = \mathrm{x}'_j\beta_i$ is estimated as $\hat{\mu}_j = \mathrm{x}'_j\hat{\beta}_i = \sum_{k\in s} \phi_{kj} y_k$ and $\sigma_j^2 \equiv \sigma^2$, the bias and variance in (6.7) are estimated as,

$$
\hat{Bias}_i = (\sum_{h=1}^{m} \sum_{j\in s_h} w_{ij}\hat{\mu}_j) - N_i^{-1}\sum_{j\in U_i} \hat{\mu}_j,\ \hat{Var}_i = N_i^{-2}\sum_{j\in s} [a_{ij}^2 + (N_i - n_i)n_i^{-1}]\lambda_j^{-1}(y_j - \hat{\mu}_j)^2,
\tag{6.8}
$$

where $\lambda_j = (1-\phi_{jj})^2 + \sum_{k\in s(-j)} \phi_{kj}^2$ and $s(-j)$ defines the sample without unit $j$.

The authors apply the procedure for estimating the PMSE of the EBLUP and the MBDE estimator (4.11) under the model (6.1), and for estimating the PMSE of the M-quantile estimator defined in Section 6.6. The first two cases condition on the model variance estimators so that the PMSE estimators do not have bias of desired order even under correct model specification. On the other hand, the estimators are shown empirically to have smaller bias than the traditional PMSE estimators in the presence of outlying observations, although with larger MSEs than the traditional estimators in the case of small area sample sizes.

### 6.2 Computation of Prediction Intervals

As in other statistical applications, very often analysts are interested in prediction intervals for the unknown area characteristics. Construction of prediction intervals under the Bayesian approach, known as *credibility intervals*, is straightforward via the posterior distribution of the predictor. A 'natural' prediction interval under the frequentist approach with desired

coverage rate $(1-\alpha)$ is $\hat{\theta}_i^{(\cdot)} \pm z_{\alpha/2}[\hat{Var}(\hat{\theta}_i^{(\cdot)} - \theta_i)]^{1/2}$, where $\hat{\theta}_i^{(\cdot)}$ is the EB, EBP or EBLUP predictor and $\hat{Var}(\hat{\theta}_i^{(\cdot)} - \theta_i)$ is an appropriate estimate of the prediction error variance. However, even under asymptotic normality of the prediction error, the use of this prediction interval has coverage error of order $O(1/m)$, which is not sufficiently accurate. Recent work in SAE focuses therefore on reducing the coverage error via parametric bootstrap.

Hall and Maiti (2006) consider the following general model: for a suitable smooth function $f_i(\beta)$ of the covariates $X_i = (X_{i1}, ..., X_{in_i})$ in area $i$ and a vector parameter $\beta$, random variables $\Theta_i = f_i(\beta) + u_i;\ E(u_i) = 0$ are drawn from a distribution $Q\{f_i(\beta); \xi\}$. The outcome observations $y_{ij}$ are drawn independently from a distribution $R\{l(\Theta_i); \eta_i\}$, where $l(\cdot)$ is a known link function and $\eta_i$ is either known or is the same for every area $i$. For given covariates $X_{i0}$, sample size $n_{i0}$ and known parameters, a $(1-\alpha)$-level prediction interval for the corresponding realization $\Theta_{i0}$ is,

(6.9) $$I_\alpha(\beta, \xi) = [q_{(1-\alpha)/2}(\beta, \xi), q_{(1+\alpha)/2}(\beta, \xi)],$$

where $q_\alpha(\beta, \xi)$ defines the $\alpha$-level quantile of the distribution $Q\{f_i(\beta); \xi\}$. A naive prediction interval with estimated parameters is $\hat{I}_\alpha(\hat{\beta}, \hat{\xi})$, but this interval has coverage error of order $O(1/m)$ and it does not use the area-specific outcome values. To reduce the error, $\hat{I}_\alpha(\hat{\beta}, \hat{\xi})$ is calibrated on $\alpha$. This is implemented by generating parametric bootstrap samples and re-estimating $\beta$ and $\xi$ similarly to the first step of the double-bootstrap procedure for PMSE estimation described in Section 6.1. Denote by $\hat{I}_\alpha^* = I_\alpha(\hat{\beta}^*, \hat{\xi}^*)$ the bootstrap interval and let $\hat{\alpha}$ denote the solution of the equation $\Pr(\theta_i^* \in \hat{I}_{\hat{\alpha}}^*) = \alpha$, where $\theta_i^* \sim Q\{f_i(\hat{\beta}), \hat{\xi}\}$. The bootstrap-calibrated prediction interval with coverage error of order $O(m^{-2})$ is $\hat{I}_{\hat{\alpha}}(\hat{\beta}, \hat{\xi})$.

Chatterjee *et al*. (2008) consider the general linear mixed model of Das *et al*. (2004) mentioned in Section 6.1; $Y = X\beta + Zu + e$, where $Y$ (of dimension $n$) signifies all the observations in all the areas, $X_{n\times p}$ and $Z_{n\times q}$ are known matrices and $u$ and $e$ are independent vector normal errors of random effects and residual terms with variance matrices $D(\psi)$ and $R(\psi)$, which are functions of a $k$-vector parameter $\psi$. Note that this model and the model of Hall and Maiti (2006) include as special cases the mixed linear models defined by (5.1) and (5.3). The present model cannot handle nonlinear mixed effects (for example, the GLMM

(5.5)), which the Hall and Maiti model can, but it does not require conditional independence of the observations given the random effects, as under the Hall and Maiti model.

The (parametric bootstrap) prediction interval of Chatterjee *et al*. (2008) for a univariate linear combination $t = c'(X\beta + Zu)$ is obtained by the following steps. First compute the conditional mean, $\mu_t$ and variance $\sigma_t^2$ of $t \mid Y; \beta, \psi$. Next generate new observations $Y^* = X\hat{\beta} + Zu^* + e^*$, where $u^* \sim N(0, D(\hat{\psi}))$, $e^* \sim N(0, R(\hat{\psi}))$. From $Y^*$ compute $\hat{\beta}^*$ and $\hat{\psi}^*$ using the same method as for $\hat{\beta}$ and $\hat{\psi}$, and compute $\hat{\mu}_t^*$ and $\hat{\sigma}_t^*$ (same as $\mu_t$ and $\sigma_t$ but with estimated parameters). Denote by $L_n^*$ the bootstrap distribution of $\hat{\sigma}_t^{-1*}(t^* - \hat{\mu}_t^*)$ where $t^* = c'(X\hat{\beta} + Zu^*)$, and let $d = (p + k)$ be the total number of unknown parameters. Then, if $d^2 / n \to 0$ and under some regularity conditions, if $q_1, q_2$ satisfy $L_n^*(q_2) - L_n^*(q_1) = 1 - \alpha$,

(6.10) $$\Pr(\hat{\mu}_t + q_1\hat{\sigma}_t \le t \le \hat{\mu}_t + q_2\hat{\sigma}_t) = 1 - \alpha + O(d^3 n^{-3/2}).$$

Note that this theory allows $d$ to grow with $n$ and that the coverage error is defined in terms of $n$ rather than $m$, the number of sampled areas, as under the Hall and Maiti (2006) approach. The total sample size increases also as the sample sizes within the areas increase, and not just by increasing $m$. By appropriate choice of $t$, the interval (6.8) is area specific.

Remark 4. The article by Chatterjee *et al*. (2008) contains a thorough review of many other prediction intervals proposed in the literature.

### 6.3 Benchmarking

Model-based SAE depends on models that can be hard to validate and if the model is misspecified, the resulting predictors may perform poorly. Benchmarking robustifies the inference by forcing the model-based predictors to agree with a design-based estimator for an aggregation of the areas for which the design-based estimator is reliable. Assuming that the aggregation contains all the areas, the benchmarking equation takes the general form,

(6.11) $$\sum_{i=1}^{m} b_i \hat{\theta}_{i,model} = \sum_{i=1}^{m} b_i \hat{\theta}_{i,design}.$$

The coefficients $\{b_i\}$ are fixed weights, assumed without loss of generality to sum to 1 (e.g., relative area sizes). The constraint (6.9) has the further advantage of guaranteeing consistency of publication between the model-based small area predictors and the design-based estimator for the aggregated area, which is often required by statistical bureaus. For example, the model-based predictors of total unemployment in counties should add up to the design-based estimate of total unemployment in the country, which is deemed accurate.

A benchmarking method in common use, often referred to as ratio or pro-rata adjustment is,

(6.12) $$\hat{\theta}_i^{bench} = (\sum_{i=1}^{m} b_i \hat{\theta}_{i,design} / \sum_{i=1}^{m} b_i \hat{\theta}_{i,model}) \times \hat{\theta}_{i,model} .$$

The use of this procedure, however, applies the same ratio correction for all the areas, irrespective of the precision of the model-based predictors before benchmarking. As a result, the prorated predictor in a given area is not consistent as the sample size in that area increases, and estimation of the PMSE of the prorated predictors is not straightforward. Consequently, other procedures have been proposed in the literature.

Wang *et al*. (2008) derive benchmarked BLUP (BBLUP) under the area level model (5.1) as the predictors minimizing $\sum_{i=1}^{m} \varphi_i E(\theta_i - \hat{\theta}_i^{bench})^2$ subject to (6.11), where the $\varphi_i$'s are chosen positive weights. The BBLUP is,

(6.13) $$\hat{\theta}_{i,BLUP}^{bench} = \hat{\theta}_{i,model}^{BLUP} + \delta_i \sum_{j=1}^{m} b_j (\theta_{j,design} - \hat{\theta}_{j,model}^{BLUP}); \ \delta_i = (\sum_{j=1}^{m} \varphi_j^{-1} b_j^2)^{-1} \varphi_i^{-1} b_i .$$

When the variance $\sigma_u^2$ is unknown, it is replaced by its estimator everywhere in (6.13), yielding the empirical BBLUP. You & Rao (2002) achieve "automatic benchmarking" for the unit level model (5.3) by changing the estimator of $\beta$. Wang *et al*. (2008) consider a similar procedure for the area level model. The latter approach is further extended by augmenting the covariates $\mathrm{x}_i$ to $\tilde{\mathrm{x}}_i' = [\mathrm{x}_i', b_i \sigma_{Di}^2]$. (The variances $\sigma_{Di}^2$ are considered known under the area level model.) The use of the augmented model yields a BLUP that likewise satisfies the benchmark constraint (6.11) and is more robust to misspecification of the mean $\mathrm{E}(y_i) = \mathrm{x}_i'\beta$ assumed under the model in terms of reducing the bias of the small area predictor.

Pfeffermann and Tiller (2006) add monthly benchmark constraints of the form (6.11) to the measurement (observation) equation of a time series state-space model fitted jointly to the direct estimates in several areas. Adding benchmark constraints to time series models is particularly important since time series models are slow to adapt to abrupt changes. The benchmarked predictor obtained under the augmented time series model belongs to the family of predictors (6.13) proposed by Wang *et. al*. (2008). By adding the constraints to the model equations, the use of this approach permits estimating the variance of the benchmarked estimators as part of the model fitting. The variance accounts for the variances of the model error terms, the variances and autocovariances of the sampling errors of the direct estimators and of the benchmarks $\sum_{i=1}^{m} b_i \hat{\theta}_{ti,direct}$, $t = 1, 2, \ldots$, and the cross-covariances and autocovariances between the sampling errors of the direct estimators and the benchmark.

Datta *et al*. (2011) develop Bayesian benchmarking by minimizing,

(6.14) $\sum_{i=1}^{m} \varphi_i E[(\theta_i - \hat{\theta}_i^{bench})^2 \mid \hat{\theta}_{design}]$ s.t. $\sum_{i=1}^{m} b_i \hat{\theta}_i^{bench} = \sum_{i=1}^{m} b_i \hat{\theta}_{i,design}$,

where $\hat{\theta}_{design} = (\hat{\theta}_{1,design},...,\hat{\theta}_{m,design})'$. The solution of this minimization problem is the same as (6.13), but with $\hat{\theta}_{k,model}^{BLUP}$ replaced everywhere by the posterior mean $\hat{\theta}_{k,Bayes}$. Denote the resulting predictors by $\hat{\theta}_{i,Bayes}^{bench,1}$. The use of these predictors has the drawback of 'over shrinkage' in the sense that $\sum_{i=1}^{m} b_i (\hat{\theta}_{i,Bayes}^{bench,1} - \bar{\hat{\theta}}_{b,Bayes}^{bench,1})^2 < \sum_{i=1}^{m} b_i E[(\theta_i - \bar{\theta}_b)^2 \mid \hat{\theta}_{design}]$, where $\bar{\hat{\theta}}_{b,Bayes}^{bench,1} = \sum_{i=1}^{m} b_i \hat{\theta}_{i,Bayes}^{bench,1}$ and $\bar{\theta}_b = \sum_{i=1}^{m} b_i \theta_i$. To deal with this problem, Datta *et al*. (2011) propose to consider instead the predictors $\hat{\theta}_{i,Bayes}^{bench,2}$ satisfying the constraints,

(6.15) $\sum_{i=1}^{m} b_i \hat{\theta}_{i,Bayes}^{bench,2} = \sum_{i=1}^{m} b_i \hat{\theta}_{i,design}$ ; $\sum_{i=1}^{m} b_i (\hat{\theta}_{i,Bayes}^{bench,2} - \sum_{i=1}^{m} b_i \hat{\theta}_{i,design})^2 = H$,

where $H = \sum_{i=1}^{m} b_i E(\theta_i - \bar{\theta}_b)^2 \mid \hat{\theta}_{design}]$. The benchmarked predictors have now the form,

(6.16) $\hat{\theta}_{i,Bayes}^{bench,2} = \sum_{i=1}^{m} b_i \hat{\theta}_{i,design} + A_{CB}(\hat{\theta}_{i,Bayes} - \bar{\hat{\theta}}_{Bayes})$ ; $A_{CB}^2 = H / \sum_{i=1}^{m} b_i (\hat{\theta}_{i,Bayes} - \bar{\hat{\theta}}_{Bayes})^2$.

Notice that the development of the Bayesian benchmarked predictors is general and not restricted to any particular model. The PMSE of the benchmarked predictor can be estimated as $E[(\hat{\theta}_{i,Bayes}^{bench,2} - \theta_i)^2 \mid \hat{\theta}_{design}] = Var(\hat{\theta}_{i,Bayes} \mid \hat{\theta}_{design}) + (\hat{\theta}_{i,Bayes}^{bench,2} - \hat{\theta}_{i,Bayes})^2$, noting that the cross-product $E[(\hat{\theta}_{i,Bayes}^{bench,2} - \hat{\theta}_{i,Bayes})(\hat{\theta}_{i,Bayes} - \theta_i) \mid \theta_{design}] = 0$.

Nandram and Sayit (2011) likewise consider Bayesian benchmarking, focusing on estimation of area proportions. Denoting by $c_i$ the number of sample units in area $i$ having characteristic $C$ and by $p_i$ the probability to have this characteristic, the authors assume the beta-binomial hierarchical Bayesian model,

(6.17)
$$c_i \mid p_i \sim Binomial(n_i, p_i);\ p_i \mid \mu, \tau \sim Beta[\mu\tau, (1-\mu)\tau],\ i = 1,...m$$
$$p(\mu, \tau) = (1+\tau^2)^{-1},\ 0 < \mu < 1,\ \tau \geq 0.$$

Let $\tilde{b}_i = n_i / n$. The benchmark constraint, called a "restriction" in the paper is defined as,

(6.18) $\sum_{i=1}^{m} \tilde{b}_i p_i = \theta;\ \theta \sim Beta[\mu_0\tau_0, (1-\mu_0)\tau_0]$.

The authors derive the joint posterior distribution of the true probabilities $\{p_i, i = 1,...,m\}$ under the unrestricted model (6.17), and the restricted model with (6.18), and prove that it is proper. Computational details are given. Different scenarios are considered regarding the prior distribution of $\theta$. Under the first scenario $\tau_0 \to \infty$, implying that $\theta$ is a point mass at $\mu_0$, assumed to be known (not a likely scenario in practice) . Under a second scenario $\mu_0$ and

$\tau_0$ are specified by the analyst. In a third scenario $\mu_0 = 0.5, \tau_0 = 2$, implying $\theta \sim Uniform(0,1)$ (non-informative prior). Theoretical arguments and empirical results show that the largest gain from using the restricted model is under the first scenario where $\theta$ is completely specified, followed by the second scenario with $\tau_0 >> 2$. No gain in precision occurs under the third scenario with a non-informative prior.

To complete this section I mention a different frequentist benchmarking procedure applied by Ugarte *et al*. (2009). By this prodedure, the small area predictors in sampled and nonsampled areas under the unit level model (5.3) are benchmarked to a synthetic estimator for a region composed of the areas as obtained under a linear regression model with heterogenehous variances (but no random effects). The benchmarked predictors are the GLS predictors of the area means under the model (5.3) among all the predictors satisfying the benchmark constraint. Notice that the GLS predictors without the constraint are the optimal predictors (5.4). For known variances the benchmarked predictors are linear, but in practice the variances are replaced by sample estimates. The authors estimate the PMSE of the resulting empirical benchmarked predictors by a single-step parametric bootstrap procedure.

### 6.4 Accounting for Measurement Errors in the Covariates

Ybarra and Lohr (2008) consider the case where some or all the covariates $\mathrm{x}_i$ in the area level model (5.1) are unknown and one uses an estimator $\hat{\mathrm{x}}_i$ obtained from another independent survey, with $MSE_D(\hat{\mathrm{x}}_i) = C_i$ under the sampling design. (For known covariates $\mathrm{x}_{ki}$, $C_{ki} = 0$.) Denoting the resulting predictor by $\hat{\theta}_i^{Err}$, it follows that for known parameters,

(6.19) $$PMSE(\hat{\theta}_i^{Err}) = PMSE(\hat{\theta}_i) + (1-\gamma_i)^2 \beta' C_i \beta ,$$

where $PMSE(\hat{\theta}_i)$ is the PMSE if one knew the true $\mathrm{x}_i$. Thus, reporting $PMSE(\hat{\theta}_i)$ in this case results in under-reporting the true PMSE. Moreover, if $\beta' C_i \beta > \sigma_u^2 + \sigma_{Di}^2$, $MSE(\hat{\theta}_i^{Err}) > \sigma_{Di}^2 = Var(\tilde{y}_i)$. The authors propose therefore to use instead the predictor,

(6.20) $$\hat{\theta}_i^{Me} = \tilde{\gamma}_i \tilde{y}_i + (1-\tilde{\gamma}_i)\hat{\mathrm{x}}_i'\beta \; ; \;\; \tilde{\gamma}_i = (\sigma_u^2 + \beta' C_i \beta) / (\sigma_{Di}^2 + \sigma_u^2 + \beta' C_i \beta) .$$

The predictor $\hat{\theta}_i^{Me}$ minimizes the MSE of linear combinations of $\tilde{y}_i$ and $\hat{\mathrm{x}}_i'\beta$. Additionally, $E(\hat{\theta}_i^{Me} - \theta_i) = (1-\tilde{\gamma}_i)[E_D(\hat{\mathrm{x}}_i) - \mathrm{x}_i]'\beta$, implying that the bias vanishes if $\hat{\mathrm{x}}_i$ is unbiased for $\mathrm{x}_i$, and $E(\hat{\theta}_i^{Me} - \theta_i)^2 = \tilde{\gamma}_i \sigma_{Di}^2 \le \sigma_{Di}^2$. The authors develop estimators for the unknown parameters $\sigma_u^2$ and $\beta$, which are then substituted in (6.16) to obtain the corresponding empirical

predictor. The PMSE of the empirical predictor is estimated using the jackknife procedure of Jiang *et al*. (2002) described in Section 6.1.

Ghosh *et al*. (2006) and Torabi *et al*. (2009) study a different situation of measurement errors. The authors assume that the true model is the unit level model (5.3) with a single covariate $x_i$ for all the units in the same area, but $x_i$ is not observed and instead different measurements $x_{ij}$ are obtained for different sampled units $j \in s_i$. The sample consist therefore of the observations $\{y_{ij}, x_{ij}; i = 1,...,m,\ j = 1,...,n_i\}$. An example giving rise to such a scenario is where $x_i$ defines the true level of air pollution in the area and the $x_{ij}$'s represent pollution measures at different sites in the area. It is assumed that $x_{ij} = x_i + \eta_{ij}$; $x_i \sim N(\mu_x, \sigma_x^2)$, and $(u_i, \varepsilon_{ij}, \eta_{ij})$ are independent normally distributed random errors with zero means and variances $\sigma_u^2, \sigma_\varepsilon^2$ and $\sigma_\eta^2$ respectively. Since $x_i$ is random, this kind of measurement error is called *structural measurement error*. The difference between the two articles is that Ghosh *et al*. (2006) only use the observations $\{y_{ij}\}$ for predicting the true area means $\overline{Y}_i$, whereas Torabi *et al*. (2009) also use the sample observations $\{x_{ij}\}$.

Assuming that all the model parameters are known, the posterior distribution of the unobserved *y*-values in area $i$ is multivariate normal, which under the approach of Torabi *et al*. (2009) yields the following Bayes predictor (also BLUP) for $\overline{Y}_i$:

$$\hat{\overline{Y}}_i^B = E(\overline{Y}_i \mid \{y_{ij}\}, \{x_{ij}\}) = (1 - f_i A_i)\overline{y}_i + f_i A_i(\beta_0 + \beta_1 \mu_x) + f_i A_i \gamma_{xi} \beta_1 (\overline{x}_i - \mu_x), \tag{6.21}$$

where $f_i = 1 - (n_i / N_i)$, $\gamma_{xi} = n_i \sigma_x^2 (\sigma_\eta^2 + n_i \sigma_x^2)^{-1}$ and $A_i = [n_i \beta_1^2 \sigma_x^2 \sigma_\eta^2 + (n_i \sigma_u^2 + \sigma_\varepsilon^2) v_i]^{-1} \sigma_\varepsilon^2 v_i$, with $v_i = (\sigma_\eta^2 + n_i \sigma_x^2)$. For large $N_i$ and small $(n_i / N_i)$, the PMSE of $\hat{\overline{Y}}_i^B$ is $E[(\hat{\overline{Y}}_i^B - \overline{Y}_i)^2 \mid \{y_{ij}, x_{ij}\}] = A_i[\beta_1^2 \sigma_x^2 + \sigma_u^2 - n_i \beta_1^2 \sigma_x^4 v_i^{-1}]$. Estimating the model parameters $\psi = (\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_u^2, \sigma_\varepsilon^2)$ by a method of moments (MOM) proposed by Ghosh *et al*. (2006) and replacing them by their estimates yields the EB estimator, which is shown to be asymptotically optimal in the sense that $m^{-1} \sum_{i=1}^m E(\hat{\overline{Y}}_i^{EB} - \hat{\overline{Y}}_i^B)^2 \to 0$ as $m \to \infty$. The PMSE of the EB predictor is estimated by a weighted jackknife procedure of Chen and Lahiri (2002).

The Bayes predictor of Ghosh *et al*. (2006) has a similar structure to (6.21) but without the correction term $f_i A_i \gamma_{xi} \beta_1 (\overline{x}_i - \mu_x)$ and with the shrinkage coefficient $A_i$ replaced by $\tilde{A}_i = [n_i(\beta_1^2 \sigma_x^2 + \sigma_u^2) + \sigma_\varepsilon^2]^{-1} \sigma_\varepsilon^2$ in the other two terms. As noted above, the authors develop a

MOM for estimating the unknown model parameters to obtain the EB predictor and prove its asymptotic optimality. They also develop an HB predictor with appropriate priors for all the parameters. The Bayes expectation and PMSE are obtained by MCMC simulations.

Ghosh and Sinha (2007) consider the same unit level model as above with sample observations $(\{y_{ij}\},\{x_{ij}\})$, but assume that the true covariate $x_i$ is a fixed unknown parameter, which is known as *functional measurement error*. The work by Ybarra and Lohr (2008) reviewed before also assumes a functional measurement error, but considers the area level model. For known parameters and $x_i$, the Bayes predictor takes now the simple form,

(6.22) $$\hat{\bar{Y}}_i^B = E(\bar{Y}_i \mid \{y_{ij}\}) = (1-f_iB_i)\bar{y}_i + f_iB_i(\beta_0+\beta_1 x_i);\ B_i = (n_i\sigma_u^2+\sigma_\varepsilon^2)^{-1}\sigma_\varepsilon^2.$$

A pseudo-Bayes predictor (PB) is obtained by substituting the sample mean $\bar{x}_i$ for $x_i$ in (6.22). A pseudo-empirical Bayes predictor (PEB) is obtained by estimating all the other unknown model parameters by the MOM developed in Ghosh *et al*. (2006). The authors show the asymptotic optimality of the PEB; $m^{-1}\sum_{i=1}^{m} E(\bar{Y}_i^{PEB}-\bar{Y}_i^{PB})^2 \to 0$ as $m\to\infty$.

Datta *et al*. (2010) propose to replace the estimator $\bar{x}_i$ of $x_i$ by its maximum likelihood estimator (MLE) under the model. The corresponding PB of $\bar{Y}_i$ (assuming known other model parameters) is the same as the PB of Ghosh and Sinha (2007), but with $B_i$ replaced by $\tilde{B}_i = (n_i\sigma_u^2+\sigma_\varepsilon^2+\beta_1^2\sigma_\eta^2)^{-1}\sigma_\varepsilon^2$. A PEB predictor is obtained by replacing the model parameters by the MOM estimators developed in Ghosh *et al*. (2006), and is shown to be asymptotically optimal using the same optimality criterion as before. The PMSE of the PEB is estimated by the Jackknife procedures of Jiang *et al*. (2002) described in Section 6.1 and the weighted jackknife procedure of Chen and Lahiri (2002). The authors report the results of a simulation study showing that their PEB predictor outperforms the PEB of Ghosh and Sinha (2007) in terms of PMSE. A modification to the predictor of Ybarra and Lohr (2008) is also proposed.

### 6.5 Treatment of Outliers

Bell and Huang (2006) consider the area level model (5.1) from a Bayesian perspective, but assume that the random effect or the sampling error (but not both) have a Student $t_{(k)}$ distribution. The *t* distribution is often used in statistical modeling to account for possible outliers because of its long tails. One of the models considered by the authors is,

(6.23) $$u_i \mid \delta_i,\sigma_u^2 \sim N(0,\delta_i\sigma_u^2);\ \delta_i^{-1} \sim Gamma[k/2,(k-2)/2],\ \ e_i \sim N(0,\sigma_{Di}^2),$$

which implies $E(\delta_i)=1$ and $u_i \mid \sigma_u^2 \sim t_{(k)}(0,\sigma_u^2(k-2)/k)$. The coefficient $\delta_i$ is distributed around 1, inflating or deflating the variance of $u_i=\theta_i-\mathrm{x}_i'\beta$. A large value $\delta_i$ signals the existence of an outlying area mean $\theta_i$. The degrees of freedom parameter, *k*, is taken as known. Setting $k=\infty$ is equivalent to assuming the model (5.1). The authors consider several possible (small) values for *k* in their application, but the choice of an appropriate value depends on data exploration. Alternatively, the authors assume the model (6.23) for the sampling error $e_i$, (with $\sigma_{Di}^2$ instead of $\sigma_u^2$), in which case it is assumed that $u_i \sim N(0,\sigma_u^2)$. The effect of assuming the model for the random effects is to push the small area predictor (the posterior mean) towards the direct estimator, whereas the effect of assuming the model for the sampling errors is to push the predictor towards the synthetic part. The use of either model is shown empirically to perform well in identifying outlying areas, but at present it is not clear how to choose between the two models. Huang and Bell (2006) extend the approach to a bivariate area level model where two direct estimates are available for every area, with uncorrelated sampling errors but correlated random effects. This model handles a situation where estimates are obtained from two different surveys.

Ghosh *et al*. (2008) likewise consider the model (5.1) and follow the EB approach. The starting point in this study is that an outlying direct estimate may arise either from a large sampling error or from an outlying random effect. The authors propose therefore to replace the EB predictor obtained from (5.2) by the robust EB predictor,

(6.24) $$\hat{\theta}_i^{Rob}=\tilde{y}_i-(1-\hat{\gamma}_i)\hat{V}_i\Psi_G[(\tilde{y}_i-\mathrm{x}_i'\hat{\beta}_{G\hat{L}S})\hat{V}_i^{-1}]\ ;\ \hat{V}_i^2=\hat{Var}(\tilde{y}_i-\mathrm{x}_i'\hat{\beta}_{GLS}),$$

where $\hat{\beta}_{G\hat{L}S}$ is the empirical GLS under the model with estimated variance $\hat{\sigma}_u^2$, and $\Psi_G$ is the Huber influence function $\Psi_G(t)=sign(t)\min(G,|t|)$ for some value $G>0$. Thus, for large positive standardized residuals $(\tilde{y}_i-\mathrm{x}_i'\hat{\beta}_{G\hat{L}S})\hat{V}_i^{-1}$, the EB $\hat{\theta}_i=\tilde{y}_i-(1-\hat{\gamma}_i)\hat{V}_i(\tilde{y}_i\text{-}\mathrm{x}_i'\hat{\beta}_{GLS})\hat{V}_i^{-1}$ under the model is replaced by $\hat{\theta}_i^{Rob}=\tilde{y}_i-(1-\hat{\gamma}_i)\hat{V}_iG$, and similarly for large negative standardized residuals, whereas in other cases the ordinary EB, $\hat{\theta}_i$ is unchanged. The value *G* may change from one area to the other and it is chosen adaptively in such a way that the excess Bayes risk under the model (5.1) from using the predictor (6.24) is bounded by some percentage point. Alternatively, *G* may be set to some constant $1\le G_0\le 2$ as often found in the robustness literature. The authors derive the PMSE of $\hat{\theta}_i^{Rob}$ under the model (5.1) for the

case where $\sigma_u^2$ is estimated by MLE with bias of order $o(1/m)$, and develop an estimator for the PMSE that is correct up to the order $O_p(1/m)$.

Under the approach of Ghosh *et al*. (2008), the EB predictor is replaced by the robust predictor (6.24), but the estimation of the unknown model parameters and the development of the PMSE and its estimator are under the original model (5.1) without accounting for possible outliers. Sinha and Rao (2009) propose to robustify also the estimation of the model parameters. The authors consider the mixed linear model (6.1), which when written compactly for all the observations $y = (y_1', ..., y_m')'$ has the form,

(6.25) $$y = X\beta + Zu + e,\ E(u) = 0,\ E(uu') = Q;\ \ E(e) = 0,\ \ E(ee') = R,$$

where $u$ is the vector of random effects and $e$ is the vector of residuals or sampling errors. The matrices $Q$ and $R$ are block diagonal with elements that are functions of a vector parameter $\zeta = (\zeta_1, ..., \zeta_L)$ of variance components such that $V(y) = V = ZQZ' + R = V(\zeta)$. The target is to predict the linear combination $\tau = l'\beta + h'u$ by $\hat{\tau} = l'\hat{\beta} + h'\hat{u}$. Under the model, the MLE of $\beta$ and $\zeta$ are obtained by solving the normal equations $X'V^{-1}(y - X\beta) = 0$ ; $(y - X\beta)'V^{-1}\frac{\partial V}{\partial \zeta_l}V^{-1}(y - X\beta) - tr(V^{-1}\frac{\partial V}{\partial \zeta_l}) = 0,\ l = 1, ..., L$. To account with possible outliers, the authors propose solving instead,

(6.26) $$X'V^{-1}U^{1/2}\Psi_G(r) = 0;\ \Psi_G'(r)U^{1/2}V^{-1}\frac{\partial V}{\partial \zeta_l}V^{-1}U^{1/2}\Psi_G(r) - tr(V^{-1}\frac{\partial V}{\partial \zeta_l}c\mathrm{I}_n) = 0,\ l = 1, ..., L,$$

where $r = U^{-1/2}(y - X\beta)$, $U = Diag[V]$, $\Psi_G(r) = [\Psi_G(r_1), \Psi_G(r_2), ...]'$ with $\Psi_G(r_k)$ defining the Huber influence function, $\mathrm{I}_n$ is the identity matrix of order $n$ and $c = E[\Psi_G^2(r_k)]$ ($r_k \sim N(0,1)$). Notice that since $Q$ and $R$ are block diagonal, the normal equations and the robust estimating equations can be written as sums over the *m* areas.

Denote by $\hat{\beta}_{Rob}, \hat{\zeta}_{Rob}$ the solutions of (6.22). The random effects are predicted by solving,

(6.27) $$Z'\hat{R}^{-1/2}\Psi_G[\hat{R}^{-1/2}(y - X\hat{\beta}_{Rob} - Zu)] - \hat{Q}^{-1/2}\Psi_G(\hat{Q}^{-1/2}u) = 0,$$

where $\hat{R} = R(\hat{\zeta}_{Rob}), \hat{Q} = Q(\hat{\zeta}_{Rob})$. Sinha and Rao (2009) estimate the PMSE of the robust small area predictors by the first step estimator in the double-bootstrap procedure of Hall and Maiti (2006) (Equation 6.4). The parameter estimates and the predictors of the random effects needed for the application of the bootstrap procedure are computed by the robust estimating equations (6.26)-(6.27), but the generation of the bootstrap samples is under the original

model with no outliers. The estimation of the PMSE can possibly be improved by generating some outlying observations, thus reflecting more closely the properties of the original sample.

### 6.6. Different Models and Estimators for Further Robustification

In this section I review four different approaches proposed in the literature for further robustification of the inference by relaxing some of the model assumptions or using different estimators. All four studies focus on the commonly used area-level and/or unit-level models defined by (5.1) and (5.3) respectively.

***M-quantile Estimation*:** Classical SAE methods model the expectations $E(y_i \mid \mathrm{x}_i, \mathrm{u}_i)$ and $E(u_i)$. Chambers and Tzavidis (2006) and Tzavidis *et al*. (2010) propose modeling instead the quantiles of the distribution $f(y_i \mid \mathrm{x}_i)$, where for now $y_i$ is a scalar. Assuming a linear model for the quantiles, this leads to a family of models indexed by the coefficient $q \in (0,1)$; $q = \Pr[y_i \le \mathrm{x}_i'\beta_q]$. In quantile regression the vector $\beta_q$ is estimated by minimizing,

$$\min_{\beta_q} \sum_{i=1}^{n} \{ | y_i - \mathrm{x}_i'\beta_q | [(1-q)\mathrm{I}(y_i - \mathrm{x}_i'\beta_q \le 0) + q\mathrm{I}(y_i - \mathrm{x}_i'\beta_q > 0)] \}. \tag{6.28}$$

*M-quantile* regression uses influence functions for estimating $\beta_q$ by solving the equations,

$$\sum_{i=1}^{n} \Psi_q(r_{iq})\mathrm{x}_i = 0;\ r_{iq} = (y_i - \mathrm{x}_i'\beta_q),\ \Psi_q(r_{iq}) = 2\Psi(s^{-1}r_{iq})[(1-q)\mathrm{I}(r_{iq} \le 0) + q\mathrm{I}(r_{iq} > 0)], \tag{6.29}$$

where $s$ is a robust estimate of scale, and $\Psi$ is an appropriate influence function. The (unique) solution $\hat{\beta}_q$ of (6.29) is obtained by an iterative reweighted least square algorithm. Note that each sample value $(y_i, \mathrm{x}_i)$ lies on one and only one of the quantiles $m_q(\mathrm{x}_i) = \mathrm{x}_i'\beta_q$ (follows from the fact that the quantiles are continuous in $q$).

How is the M-quantile theory used for SAE? Suppose that the sample consists of unit level observations $\{y_{ij}, \mathrm{x}_{ij};\ i = 1,...,m,\ j = 1,...,n_i\}$. Identify for unit $(i, j)$ the value $q_{ij}$ such that $\mathrm{x}_{ij}'\hat{\beta}_{q_{ij}} = y_{ij}$. A predictor of the mean $\theta_i$ in area $i$ is obtained by averaging the quantiles $q_{ij}$ over the sampled units $j \in s_i$ and computing,

$$\hat{\theta}_i^M = N_i^{-1}\left(\sum_{j \in s_i} y_{ij} + \sum_{k \notin s_i} \mathrm{x}_{ik}'\hat{\beta}_{\bar{q}_i}\right);\ \bar{q}_i = \sum_{j=1}^{n_i} q_{ij} / n_i. \tag{6.30}$$

Alternatively, one can average the vector coefficients $\beta_{q_{ij}}$ and replace $\hat{\beta}_{\bar{q}_i}$ in (6.30) by the mean $\bar{\hat{\beta}}_i = \sum_{j=1}^{n_i} \hat{\beta}_{q_{ij}} / n_i$. The vectors $\hat{\beta}_{\bar{q}_i}$ or $\bar{\hat{\beta}}_i$ account for differences between the areas, similarly to the random effects under the unit level model (5.3).

The use of this approach is not restricted to the estimation of means although it does assume continuous *y*-values. For example, the distribution function in area $i$ can be estimated as $\hat{F}_i(t) = N_i^{-1}[\sum_{j\in s_i} \mathrm{I}(y_{ij} \le t) + \sum_{k\notin s_i} \mathrm{I}(\mathrm{x}'_{ik}\overline{\hat{\beta}}_i \le t)]$. Chambers and Tzavidis (2006) develop unconditional and area specific estimators for the variance of the M-quantile estimators assuming $\hat{\beta}_{\bar{q}_i}$ (or $\overline{\hat{\beta}}_i$) is fixed, and bias estimators under the linear model $E(y_{ij} \mid x_{ij}) = x'_{ij}\beta_i$.

The M-quantile approach does not assume a parametric model although it assumes that the quantiles are linear in the covariates in the theory mentioned above. Clearly, if the unit level model (5.3) holds, the use of the model is more efficient, but the authors illustrate that the M-quantile predictors can be more robust to model misspecification. Notice in this regard that the approach is not restricted to a specific definition of the small areas. It accounts also for possible outliers by choosing an appropriate influence function in the estimating equation (6.29). On the other hand, there seems to be no obvious way of how to predict the means or other target quantities for nonsampled areas. A possible simple solution would be to set $q = 0.5$ for such areas or weight the *q*-values of neighboring sampled areas, but it raises the question of how to estimate the corresponding PMSE, unless under a model.

***Use of Penalized Spline Regression*:** Another way of robustifying the inference is by use of penalized spline (P-spline) regression. The idea here is to avoid assuming a specific functional form for the expectation of the response variable. Suppose that there is a single covariate $x$. The P-spline model assumes $y = m_0(\mathrm{x}) + \varepsilon$, $E(\varepsilon) = 0, Var(\varepsilon) = \sigma_\varepsilon^2$. The mean $m_0(x)$ is taken as unknown and approximated as,

(6.31) $\quad m(x;\beta,\gamma) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{K} \gamma_k (x - K_k)_+^p$; $\ (x - K_k)_+^p = \max[0, (x - K_k)^p]$,

where $p$ is the degree of the spline and $K_1 < ... < K_k$ are fixed knots. For large $K$ and good spread of the knots over the range of *x*, the spline (6.31) approximates well most smooth functions. It uses the basis $[1, x, ..., x^p, (x - K_1)_+^p, ..., (x - K_k)_+^p]$ to approximate the mean, but other bases can be considered, particularly when there are more covariates.

Opsomer *et al*. (2008) use P-spline regression for SAE by treating the $\gamma$-coefficients in (6.31) as additional random effects. Suppose that the data consist of unit-level observations, $\{y_{ij}, \mathrm{x}_{ij};\ i = 1, ..., m,\ j = 1, ..., n_i\}$. For unit $j$ in area $i$, the model considered is

(6.32) $\qquad y_{ij} = \beta_0 + \beta_1 x_{ij} + ... + \beta_p x_{ij}^p + \sum_{k=1}^{K} \gamma_k (x_{ij} - K_k)_+^p + u_i + \varepsilon_{ij}$,

where the $u_i$'s are the usual area random effects and $\varepsilon_{ij}$'s are the residuals. Let $u=(u_1,...,u_m)'$, $\gamma=(\gamma_1,...,\gamma_K)'$. Defining $d_{ij}=1(0)$ if unit $j$ is (is not) in area $i$ and denoting $d_j=(d_{1j},...,d_{mj})'$ and $D=[d_1,...,d_n]'$, the model holding for the vector $y$ of all the response values can be written compactly as,

(6.33) $$y=X\beta+Z\gamma+Du+\varepsilon;\ \gamma\sim(0,\sigma_\gamma^2\mathrm{I}_k),\ u\sim(0,\sigma_u^2\mathrm{I}_m),\ \varepsilon\sim(0,\sigma_\varepsilon^2\mathrm{I}_n),$$

where $X=[x_1^{(p)},...,x_n^{(p)}]'$ with $x_l^{(p)}=(1,x_l,...,x_l^p)'$, and $Z=[z_1,...,z_n]'$ with $z_l=[(x_l-K_1)_+^p,...,(x_l-K_k)_+^p)]'$. The model (6.33) looks similar to (6.25) but the responses $y_{ij}$ are not independent between the areas because of the common random effects $\gamma$. Nonetheless, the BLUP and EBLUP of $(\beta,u,\gamma)$ can be obtained using standard results. See the article for the appropriate expressions. The small area EBLUP are obtained as,

(6.34) $$\hat\theta_{i,EBLUP}^{\text{P-}spline}=\hat\beta'\bar X_i^{(p)}+\hat\gamma'\bar Z_i+\hat u_i;\ \bar X_i^{(p)}=\textstyle\sum_{l\in U_i}x_l^{(p)}/N_i,\ \bar Z_i=\sum_{l\in U_i}z_l/N_i.$$

The use of this approach requires that the covariates are known for every element in the population. Opsomer *et al*. (2008) derive the PMSE of the EBLUP (6.34) correct to second order for the case where the unknown variances are estimated by REML, and an estimator of the PMSE with bias correct to the same order. The authors develop also a nonparametric bootstrap algorithm for estimating the PMSE and for testing the hypotheses $\sigma_u^2=0$ and $\sigma_\gamma^2=0$ of no random effects. Rao *et al*. (2009) use a similar model to (6.33) but rather than computing the EBLUP under the model, the authors propose estimators that are robust to outliers, similarly (but not the same) to the methodology developed by Sinha and Rao (2009) for the mixed linear model described in Section 6.5. Jiang *et al*. (2010) show how to select an appropriate spline model by use of the fence method. See Section 8.

***Use of Empirical Likelihood in Bayesian Inference*:** Chaudhuri and Ghosh (2011) consider the use of empirical likelihoods (EL) instead of fully parametric likelihoods as another way of robustifying the inference. When combined with appropriate proper priors, it defines a semiparametric Bayesian approach, which can handle continuous and discrete outcomes in area- and unit-level models without specifying the distribution of the outcomes as under the classical Bayesian approach. Denote as before by $\theta=(\theta_1,...,\theta_m)'$ and $y=(y_1,...,y_m)'$ the area parameters and the corresponding direct estimators, with jumps $\tau=(\tau_1,...,\tau_m)$ defining the empirical distribution of $y$, so that $\sum_{i=1}^m\tau_i=1$. The (maximum) EL is $L_E(\theta)=\prod_{i=1}^m\hat\tau_i(\theta)$,

where for given moments $E(y_i \mid \theta_i) = k(\theta_i)$, $Var(y_i \mid \theta_i) = V(\theta_i)$, the estimate $\hat{\tau}(\theta)$ is the solution of the constrained maximization problem,

(6.35) $\max_{\tau_1,...,\tau_m} \prod_{i=1}^{m} \tau_i$, *s.t.* $\tau_i \geq 0$, $\sum_{i=1}^{m} \tau_i = 1$; $\sum_{i=1}^{m} \tau_i [y_i - k(\theta_i)] = 0$, $\sum_{i=1}^{m} \tau_i \{\frac{[y_i - k(\theta_i)]^2}{V(\theta_i)} - 1\} = 0$.

Under the area model (5.1) $k(\theta_i) = \theta_i = \mathrm{x}_i'\beta + u_i$ and $V(\theta_i) = \sigma_{Di}^2$. The authors assume proper priors for $(\beta, u_1, ..., u_m, \sigma_u^2)$ and hence for $\theta$, thus guaranteeing that the posterior distribution $\pi(\theta \mid y)$ is also proper. For given $\theta$ the constrained maximization problem (6.35) is solved by standard methods (see the article) and by combining the EL with the prior distributions, observations from the posterior distribution $\pi(\theta \mid y)$ are obtained by MCMC simulations.

For the unit-level model (5.3), $E(y_{ij} \mid \theta_{ij}) = k(\theta_{ij}) = \mathrm{x}_{ij}'\beta + u_i$ and $Var(y_{ij} \mid \theta_{ij}) = V(\theta_{ij}) = \sigma_\varepsilon^2$. Denoting by $\tau_{ij}$ the jumps of the empirical distribution in area $i$, the EL is defined in this case as $L_E(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \hat{\tau}_{ij}(\theta) = \prod_{i=1}^{m} \hat{\tau}_{(i)}(\theta)$, where for given $\theta_{(i)} = (\theta_{i1}, ..., \theta_{i,n_i})'$, $\hat{\tau}_{(i)}(\theta) = [\hat{\tau}_{i1}(\theta), ..., \hat{\tau}_{in_i}(\theta)]'$ is the solution of the area specific maximization problem,

(6.36) $\max_{\{\tau_{ij}\}} \prod_{j=1}^{n_i} \tau_{ij}$, *s.t.* $\tau_{ij} \geq 0$, $\sum_{j=1}^{n_i} \tau_{ij} = 1$; $\sum_{j=1}^{n_i} \tau_{ij} [y_{ij} - k(\theta_{ij})] = 0$, $\sum_{j=1}^{n_i} \tau_{ij} \{\frac{[y_{ij} - k(\theta_{ij})]^2}{V(\theta_{ij})} - 1\} = 0$.

The authors applied the procedure for estimating state-wise median income of four-person families in the U.S.A. using the area-level model. Comparisons with the census values for the same year reveal much better predictions under the proposed approach compared to the direct survey estimates and the HB predictors obtained under normality of the direct estimates.

***Best predictive SAE*:** In the three previous approaches reviewed in this section the intended robustification is achieved by relaxing some of the model assumptions. Jiang *et al*. (2011) propose instead to change the estimation of the fixed model parameters. The idea is simple. In classical model-based SAE the EBLUP or EB predictors are obtained by replacing the parameters in the expression of the BP by their MLE or REML estimators. Noting that in SAE the actual target is the prediction of the area means and the estimation of model parameters is just an intermediate step, the authors propose to estimate the fixed parameters in such a way that the resulting predictors are optimal under some loss function.

Consider the area-level model (5.1) with normal errors and suppose first that $\sigma_u^2$ is known. Under the model, $E(y_i) = \mathrm{x}_i'\beta$ but suppose that the model is misspecified and $E(y_i) = \mu_i$, such that $\theta_i = \mu_i + u_i$, $i = 1, ..., m$. Let $\tilde{\theta}_i$ be a predictor of $\theta_i$ and define the loss function to be

$MSPE(\tilde{\theta}) = \sum_{i=1}^{m} E(\tilde{\theta}_i - \theta_i)^2$, where the expectation is under the *correct model*. By (5.2), the MSPE of the BP for given $\beta$ is $MSPE[\tilde{\theta}(\beta)] = E\{\sum_{i=1}^{m}[\gamma_i y_i + (1-\gamma_i)\mathrm{x}_i'\beta - \theta_i]^2\}$. The authors propose minimizing the expression inside the expectation with respect to $\beta$, which is shown to be equivalent to minimizing $\sum_{i=1}^{m}[(1-\gamma_i)^2(\mathrm{x}_i'\beta)^2 - 2\sum_{i=1}^{m}(1-\gamma_i)^2\mathrm{x}_i'\beta y_i]$, yielding the 'best predictive estimator' (BPE)

(6.37) $$\tilde{\beta} = [\sum_{i=1}^{m}(1-\gamma_i)^2\mathrm{x}_i\mathrm{x}_i']^{-1}\sum_{i=1}^{m}(1-\gamma_i)^2\mathrm{x}_i y_i .$$

Notice that unless $Var_D(e_i) = \sigma_{Di}^2 = \sigma_D^2$, $\tilde{\beta}$ differs from the commonly used GLS estimator under the model (5.1); $\hat{\beta}_{GLS} = [\sum_{i=1}^{m}\gamma_i\mathrm{x}_i\mathrm{x}_i']^{-1}\sum_{i=1}^{m}\gamma_i\mathrm{x}_i y_i$. The 'observed best predictor' (OBP) of $\theta_i$ is obtained by replacing $\hat{\beta}_{GLS}$ by $\tilde{\beta}$ in the BP (5.2) under the model (5.1).

The authors derive also the BPE of $\psi = (\beta', \sigma_u^2)'$ for the case where $\sigma_u^2$ is unknown, in which case the OBP is obtained by replacing $\sigma_u^2$ and $\hat{\beta}_{GLS}$ by the BPE of $\psi$ in (5.2). Another extension is for the unit level model (5.3), with the true area means and loss function defined as $\theta_i = \overline{Y}_i$ and $MSPE[\tilde{\theta}(\psi)] = \sum_{i=1}^{m} E_D[\tilde{\theta}_i(\psi) - \theta_i]^2$ respectively, where $\psi = (\beta', \sigma_u^2, \sigma_\varepsilon^2)'$ and $E_D(\cdot)$ is the design (randomization) expectation over all possible sample selections (Section 4.1). The reason for using the design expectation in this case is that it is almost free of model assumptions. Theoretical derivations and empirical studies using simulated data and a real data set illustrate that the OBP can outperform very significantly the EBLUP in terms of PMSE if the underlying model is misspecified. The two predictors are shown to have similar PMSE under correct model specification.

### 6.7 Prediction of Ordered Area Means

Malinovsky and Rinott (2010) consider the following (hard) problem: predict the ordered area means $\theta_{(1)} \le \theta_{(2)} \le ... \le \theta_{(m)}$ under the area-level model $\tilde{y}_i = \mu + u_i + e_i = \theta_i + e_i$ (special case of 5.1), with $u_i \overset{iid}{\sim} H(0, \sigma_u^2)$, $e_i \overset{iid}{\sim} G(0, \sigma_e^2)$; $H$ and $G$ are general distributions with zero means and variances $\sigma_u^2$ and $\sigma_e^2$. To illustrate the difference between the prediction of ordered and unordered means, consider the prediction of $\theta_{(m)} = \max_i\{\theta_i\}$. If $\hat{\theta}_i$ satisfies $E(\hat{\theta}_i \mid \theta_i) = \theta_i,\ i = 1, ..., m$, then $E[\max_i\{\hat{\theta}_i\} \mid \{\theta_j\}] \ge \theta_{(m)}$ so that the largest estimator

overestimates the true largest mean. On the other hand, the Bayesian predictors $\theta_i^* = E[\theta_i \mid \{\hat{\theta}_j\}]$ satisfy $E[\max_i\{\theta_i^*\}] < E(\theta_{(m)})$, an underestimation in expectation.

Wright, Stern and Cressie (2003) considered the prediction of ordered means from a Bayesian perspective but their approach requires heavy numerical calculations and is sensitive to the choice of priors. Malinovsky and Rinott (2010) compare three predictors of the ordered means under the frequentist approach using the loss function $L(\tilde{\theta}_{()}, \theta_{()}) = \sum_{i=1}^{m}(\tilde{\theta}_{(i)} - \theta_{(i)})^2$ and the Bayes risk $E[L(\tilde{\theta}_{()}, \theta_{()})]$. Let $\hat{\theta}_i$ define the direct estimator of $\theta_i$ and $\hat{\theta}_{(i)}$ the $i$-th ordered estimator (statistic). The predictors compared are:

(6.38) $\tilde{\theta}_{(i)}^{(1)} = \hat{\theta}_{(i)}$; $\tilde{\theta}_{(i)}^{(2)}(\delta) = \delta\hat{\theta}_{(i)} + (1-\delta)\bar{\hat{\theta}}$, $\bar{\hat{\theta}} = \sum_{i=1}^{m}\hat{\theta}_i / m$; $\tilde{\theta}_{(i)}^{(3)} = E(\theta_{(i)} \mid \hat{\theta})$, $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)'$.

The results below assume that $\sigma_u^2$ and $\sigma_e^2$ are known and $\mu$ is estimated by $\bar{\hat{\theta}}$.

Denote by $\tilde{\theta}^{[k]}$ the predictor of the ordered means when using the predictors $\tilde{\theta}_{(i)}^{(k)}$, $k = 1, 2, 3$ and let $\gamma = \sigma_u^2(\sigma_u^2 + \sigma_e^2)^{-1}$ be the shrinkage coefficient when predicting the unordered means (Eq. 5.2). The authors derive several theoretical comparisons. For example,

(6.39) If $\gamma \le (m-1)^2/(m+1)^2$ then $E[L(\tilde{\theta}_{()}^{[2]}(\delta), \theta_{()})] \le E[L(\tilde{\theta}_{()}^{[1]}, \theta_{()})]$ for all $\gamma \le \delta \le 1$.

Noting that $\lim_{m\to\infty}[(m-1)^2/(m+1)^2] = 1$, it follows that (6.39) holds asymptotically for all $\gamma$ and the inequality $\gamma \le \delta \le 1$ implies less shrinkage of the direct estimators towards the mean. In particular, the optimal choice of $\delta$ for $\tilde{\theta}^{[2]}(\delta)$ satisfies, $\lim_{m\to\infty}\delta^{opt} = \gamma^{1/2}$.

The results above assume general distributions *H* and *G*. When these distributions are normal, then for $m = 2$, $E[L(\tilde{\theta}_{()}^{[3]}, \theta_{()})] \le E[L(\tilde{\theta}_{()}^{[2]}(\delta), \theta_{()})]$ for all $\delta$. A conjecture supported by simulations is that this relationship holds also for $m > 2$. However, the simulations suggest that for sufficiently large $m$ (e.g., $m \ge 25$), $\tilde{\theta}^{[3]}$ is efficiently replaced by $\tilde{\theta}^{[2]}(\gamma^{1/2})$. The last two conclusions are shown empirically to hold also in the case where $\sigma_u^2$ is unknown and replaced by the MOM variance estimator.

Remark 5. The problem of predicting the ordered means is different from ranking them, one of the famous triple-goal estimation objectives in SAE. The triple-goal estimation consists of producing 'good' area specific estimates, 'good' estimates of the histogram (distribution) and 'good' estimates of the ranks. See Rao (2003) for discussion. Judkins and Liu (2000) considered another related problem of estimating the range of the area means. The authors show theoretically and by simulations that the range of the direct estimators overestimates

the true range, whereas the range of the empirical Bayes estimators underestimates the true range, in line with the discussion at the beginning of this section. The bias is much reduced by use of a constrained empirical Bayes estimator. For the model considered by Malinovsky and Rinott (2010), the constrained estimator is obtained by replacing the shrinkage factor $\gamma = \sigma_u^2(\sigma_u^2 + \sigma_e^2)^{-1}$ in (5.2) by $\tilde{\gamma} \cong \gamma^{-1/2}$, which again shrinkages less the direct estimator.

**6.8. New developments for specific applications**

In this section I review two relatively new applications of SAE; assessment of literacy and poverty mapping. The latter application, in particular, received considerable attention in recent years.

***Assessment of literacy*:** The notable feature of assessing literacy from a literacy test is that the possible outcome is either zero, indicating illiteracy, or a positive continuous score measuring the level of literacy. Another example is the consumption of illicit drug, where the consumption is either zero or a continuous measure. In both examples the zero scores are "structural" (true) zeroes. The common models used for SAE are not applicable for this kind of responses if the proportion of zeroes is high. Pfeffermann *et al*. (2008) consider the estimation of the average literacy and the proportion of people with positive scores in districts and villages in Cambodia, a study sponsored by the UNESCO Institute for Statistics(UIS). Denote by $y_{ijk}$ the test score of adult $k$ from village $j$ of district $i$ and by $r_{ijk}$ a set of covariates and district and village random effects. The following relationship holds:

(6.40) $$E(y_{ijk} \mid r_{ijk}) = E(y_{ijk} \mid r_{ijk}, y_{ijk} > 0)\Pr(y_{ijk} > 0 \mid r_{ijk}).$$

The two parts in the right hand side of (6.40) are modeled as: $E[y_{ijk} \mid r_{ijk}, y_{ijk} > 0] = \mathrm{x}'_{ijk}\beta + u_i + v_{ij}$ where $(u_i, v_{ij})$ are district and nested village random effects, $\Pr(y_{ijk} > 0 \mid r_{ij}) = p_{ijk}$; $\mathrm{logit}(p_{ijk}) = \gamma' \mathrm{z}_{ijk} + u_i^* + v_{ij}^*$ where $\mathrm{z}_{ijk}$ defines a set of covariates which may differ from $\mathrm{x}_{ijk}$ and $(u_i^*, v_{ij}^*)$ are district and nested village random effects, which are correlated respectively with $(u_i, v_{ij})$. The village and district predictors of the average score and the proportion of positive scores are obtained by application of the Bayesian approach with noninformative priors using MCMC simulations. The use of the Bayesian approach enables to account for the correlations between the respective random effects in the two models, which is not feasible when fitting the two models separately. The area predictors are obtained by imputing the responses for nonsampled individuals by sampling from their posterior distribution, which are then added to the observed responses (when responses exist).

Remark 6. Mohadjer *et al*. (2007) estimate the proportions $\theta_{ij}$ of adults in the lowest level of literacy in counties and states of the U.S.A., by modeling the direct estimates $\tilde{p}_{ij}$ in county $j$ of state $i$ as $\tilde{p}_{ij} = \theta_{ij} + \varepsilon_{ij}$, and modeling $\text{logit}(\theta_{ij}) = x'_{ij}\beta + u_i + v_{ij}$ with $u_i$ and $v_{ij}$ defining state and county random effects. The state and county estimates are likewise obtained by MCMC simulations with noninformative priors. Note that this is not a two-part model.

***Poverty mapping*:** The estimation of poverty indicators in small regions is of major concern in many countries across the world, initiated and sponsored in many cases by the United Nations and the World Bank. In a celebrated article (awarded by the Canadian Statistical Society as the best paper published in 2010 in *The Candian Journal of Statitcis*), Molina and Rao focus on estimation of area means of nonlinear poverty measures called FGT defined as,

$$F_{\alpha i} = \frac{1}{N_i}\sum\nolimits_{j=1}^{N_i} F_{\alpha ij};\; F_{\alpha ij} = \left(\frac{z - E_{ij}}{z}\right)^{\alpha} \times \mathrm{I}(E_{ij} < z),\; \alpha = 0,1,2\,, \tag{6.41}$$

where $E_{ij}$ is a measure of welfare for unit $j$ in area $i$ such as income or expenditure, $z$ is a poverty treshold under which a person is considered 'poor' (e.g., 60% of the nation median income), and $\mathrm{I}(\cdot)$ is the indicator function. For $\alpha = 0$ $F_{\alpha i}$ is the proportion under poverty. For $\alpha = 1$ $F_{\alpha i}$ measures the "poverty gap" and for $\alpha = 2$ $F_{\alpha i}$ measures "poverty severity".

For $\alpha = 1,2$ it is practically impossible to assign a distribution for the measures $F_{\alpha ij}$ and in order to estimate the means $F_{\alpha i}$ in sampled and nonsampled areas, Rao and Molina (2012) assume the existence of a one-to-one transformation $y_{ij} = T(E_{ij})$ such that the transformed outcomes $y_{ij}$ satisfy the unit level model (5.3) with normal distribution of the random effects and the residuals. Notice that $F_{\alpha ij} = [1 - \frac{1}{z}T^{-1}(y_{ij})]^{\alpha} \times \mathrm{I}(T^{-1}(y_{ij} < z) =: h_{\alpha}(y_{ij})$. For sampled unit $j \in s_i$ $F_{\alpha ij}$ is known and for the nonsampled units $k \in r_i$ the missing outcomes are predicted by the EBP $F_{\alpha ik}^{EBP} = \hat{E}[h_{\alpha}(y_{ik}) \mid y_s] = \sum_{l=1}^{L} h_{\alpha}(y_{ik}^{(l)})/L$ with large $L$, where $y_s$ defines all the observed outcomes. The predictions $y_{ik}^{(l)}$ are obtained by Monte Carlo simulation from the conditional normal distribution of the unobserved outcomes given the observed outcomes under the model (5.3), using estimated parameters $\hat{\psi} = (\hat{\beta}', \hat{\sigma}_u^2, \hat{\sigma}_{\varepsilon}^2)'$. The PMSE of the EBP $\hat{F}_{\alpha i}^{EBP} = [\sum_{j \in s_i} F_{\alpha ij} + \sum_{k \in r_i} F_{\alpha ik}^{EBP}]/N_i$ is estimated similarly to the first step of the double-bootstrap procedure described in Section 6.1. Model- and design-based

simulations and application to a real data set in Spain using the transformation $y_{ij} = \log(E_{ij})$ demonstrate good performance of the area predictors and the PMSE estimators.

Remark 7. The World Bank (WB) is currently using a different method underwich all the population outcomes are simulated from the model (5.3) with estimated parameters (including for sampled units), but with random effects for design clusters, which may be different from the small areas. As discussed and by Molina and Rao (2010), the use of this procedure means that all the areas are practically considered as nonsampled and hence the resulting predictors of the means $F_{\alpha i}$ in (6.41) are synthetic predictors since the random effects and the area means of the residuals cancel out over the $L$ simulated populations. The simulation results in Molina and Rao (2012) show that the WB method produces predictors with much larger PMSE than the PMSE of the EBP predictors proposed by them.

## 7. SAE UNDER INFORMATIVE SMPLING AND NONRESPONSE

All the studies reviewed in this paper assume at least implicitly that the selection of areas that are sampled and the sampling designs within the selected areas are noninformative, implying that the model assumed for the population values applies also to the sample data with no sampling effects. This, however, may not be the case and as illustrated in the literature, ignoring the effects of informative sampling may bias the inference quite severely. A similar problem is informative nonresponse under which the response probabilities depend on the missing data, which again can distort the predictions if not accounted for properly. These problems received attention under both the frequentist and the Bayesian approaches.

Pfeffermann and Sverchkov (2007) consider the problem of informative sampling of areas and within the areas. The basic idea in this article is to fit a sample model to the observed data and then exploit the relationship between the sample model, the population model and the sample-complement model (the model holding for nonsampled units) in order to obtain unbiased predictors for the means in sampled and nonsampled areas.

Consider a two-stage sampling design by which *m* out of *M* areas are selected in the first stage with probabilities $\pi_i = \Pr(i \in s)$, and $n_i$ out of $N_i$ units are sampled from the i$^{th}$ selected area with probabilities $\pi_{j|i} = \Pr(j \in s_i \mid i \in s)$. Denote by $\mathrm{I}_i$ and $\mathrm{I}_{ij}$ the sample indicator variables for the two sampling stages and by $w_i = 1/\pi_i$ and $w_{j|i} = 1/\pi_{j|i}$ the first and second stage sampling weights. Suppose that the first level area random effects $\{u_1,...,u_M\}$ are generated independently from a distribution with *pdf* $f_p(u_i)$, and that for given $u_i$ the

second level values $\{y_{i1},...,y_{iN_i}\}$ are generated independently from a distribution with *pdf* $f_p(y_{ij} \mid x_{ij}, u_i)$, for $j = 1,...,N_i$. The conditional first-level *sample pdf* of $u_i$ that is, the *pdf* of $u_i$ for area $i \in s$ is,

$$(7.1) \qquad f_s(u_i) \overset{def}{=} f(u_i \mid \mathrm{I}_i = 1) = \Pr(\mathrm{I}_i = 1 \mid u_i) f_p(u_i) / \Pr(\mathrm{I}_i = 1) = E_s(w_i) f_p(u_i) / E_s(w_i \mid u_i).$$

The conditional first-level *sample-complement pdf* of $u_i$, that is, the *pdf* for area $i \notin s$ is,

$$(7.2) \qquad f_c(u_i) \overset{def}{=} f(u_i \mid \mathrm{I}_i = 0) = \Pr(\mathrm{I}_i = 0 \mid u_i) f_p(u_i) / \Pr(\mathrm{I}_i = 0).$$

Note that the *population*, *sample* and *sample-complement pdf*s are the same if $\Pr(\mathrm{I}_i = 1 \mid u_i) = \Pr(\mathrm{I}_i = 1)$, in which case the area selection is *noninformative*. Similar relationships hold between the sample model, population model and sample-complement models for the outcomes $y_{ij}$ within the selected areas for given values of the random effects.

Pfeffermann and Sverchkov (2007) illustrate their procedure by assuming that the *sample model* is the unit-level model (5.3) with normal random effects and residuals, and that the sampling weights within the selected areas have sample model expectations:

$$(7.3) \qquad E_{si}(w_{j|i} \mid \mathrm{x}_{ij}, y_{ij}, u_i, \mathrm{I}_i = 1) = E_{si}(w_{j|i} \mid \mathrm{x}_{ij}, y_{ij}, \mathrm{I}_i = 1) = k_i \exp(a'\mathrm{x}_{ij} + b y_{ij}),$$

where $k_i = N_i (n_i)^{-1} \sum_{j=1}^{N_i} \exp(-a x_{ij} - b y_{ij}) / N_i$ and $a$ and $b$ are fixed constants. No model is assumed for the relationship between the area selection probabilities and the area means. The authors show that under this model and for given parameters $\{\beta, b, \sigma_u^2, \sigma_\varepsilon^2\}$, the true mean $\bar{Y}_i$ in sampled area $i$ is predicted as,

$$(7.4) \qquad \hat{\bar{Y}}_i = E_p(\bar{Y}_i \mid D_s, \mathrm{I}_i = 1) = \frac{1}{N_i}\{(N_i - n_i)\hat{\theta}_i + n_i[\bar{y}_i + (\bar{X}_i - \bar{\mathrm{x}}_i)'\beta] + (N_i - n_i) b \sigma_e^2\},$$

where $D_s$ represents all the known data and $\hat{\theta}_i = \hat{u}_i + \bar{X}_i \beta$ is the optimal predictor of the sample model mean $\theta_i = \bar{X}_i' \beta + u_i$. The last term in (7.4) corrects for the sample selection effect, that is, the difference between the sample-complement expectation and the sample expectation in sampled areas.

The mean $\bar{Y}_k$ of area $k$ not in the sample is predicted as,

$$(7.5) \qquad \hat{E}_p(\bar{Y}_k \mid D_s, \mathrm{I}_k = 0) = \bar{X}_k' \beta + b\sigma_e^2 + [\sum_{i \in s} (w_i - 1)\hat{u}_i / \sum_{i \in s} (w_i - 1)].$$

The last term of (7.5) corrects for the fact that the mean of the random effects in areas outside the sample is different from zero under informative selection of the areas. The authors develop test procedures for testing the informativeness of the sample selection and develop a

bootstrap procedure for estimating the PMSE of the empirical predictors obtained by substituting the unknown model parameters by sample estimates. The method is applied for predicting the mean body mass index (BMI) in counties of the U.S.A using data from the third national health and nutrition examination survey (NHANES III ).

Malec *et al*. (1999, hereafter MDC) and Nandram and Choi (2010, hereafter NC) likewise consider the estimation of county level BMI statistics from NHANES III, with both articles accounting for within area informative sampling in a similar manner, and the latter article accounting in addition for informative nonresponse. Another difference between the two articles is that MDC consider binary population outcomes (overweight/normal status), with logistic probabilities that contain fixed and multivariate random area effects, whereas NC assume a log-normal distribution for the continuous BMI measurement, with linear spline regressions containing fixed and random area effects, defining the means. In order to account for sampling effects, both articles assume that each sampled unit represents $K-1$ other units within a specified group (cluster) of units, with unit $j$ selected with probability $\pi^*_{(j)}$ that can take one of the $G$ observed values $\pi^*_g, g=1,...,G$ in that group. Specifically, let $\delta_j = 1(0)$ if unit $j$ is sampled (not sampled). The MDC model assumes,

$$(7.6)\quad \begin{array}{l} \delta_j \mid K, \pi^*_{(j)} \overset{ind}{\sim} \text{Bernoulli}(\pi^*_{(j)}), j=1,...,K;\ \Pr(\pi^*_{(j)} = \pi^*_g \mid \theta_{gy}, y_j = y) = \theta_{gy},\ y=0,1;\ g=1,...,G \\ \Pr(y_j = y \mid p) = p^y(1-p)^{1-y}, 0 \le p \le 1;\ p(K)=1 \end{array}.$$

It follows that,

$$(7.7)\quad \mathrm{P}(\delta_j = 1, y_j = y, \pi^*_{(j)} = \pi^*_g, \{\delta_k = 0\}_{k \ne j} \mid \theta, p) \propto \frac{p^y(1-p)^{1-y}}{\sum_{g=1}^{G} \pi^*_g \sum_{y=0}^{1} \theta_{gy} p^y (1-p)^{1-y}}.$$

MDC show that the MLE of $\theta_{gy}$ is $\hat{\theta}_{gy} = (\tau_{gy} / \pi^*_g) / \sum_{g^*=1}^{G} (\tau_{g^*y} / \pi^*_{g^*})$ where $\tau_{gy}$ is the sample frequency of $\pi^*_g$ in the group for units with overweight status $y$. They plug the estimate into (7.7) and then into the full likelihood that includes also the distribution of the random effects.

NC generalize the model (7.6) by allowing the outcome to be continuous assuming $\Pr(\pi^*_{(j)} = \pi^*_g \mid \theta_g(y), y) = \theta_g(y), -\infty < y < \infty$ where $\theta_g(y) = \theta_{gl}$ for $a_{l-1} < y < a_l$, and replacing the Bernoulli distribution by a continuous *pdf*. To account for informative nonresponse, the authors assume that the response probabilities $p^r_{ij}$ are logistic with $\text{logit}(p^r_{ij}) = v_{0i} + v_{1i} y_{ij}$, where $\{(v_{0i}, v_{1i})\}$ is another set of random effects having a bivariate normal distribution.

Remark 8. As the notation suggests, both MDC and NC use the full Bayesian approach with appropriate prior distributions to obtain the small area means under the respective models. See the articles for details. The authors do not consider informative sampling of the areas.

I conclude this section by describing an article by Zhang (2009), which uses a very different model from the other models considered in the present paper. The article considers the estimation of small area compositions in the presence of informative nonresponse. Compositions are the counts or proportions in categories of a categorical variable such as types of households, and estimates of the compositions are required for every area. Zhang deals with this problem by assuming that the generalized SPREE model (GSPREE) developed in Zhang and Chambers (2004) holds for the complete data (with no missingness). In order to account for the nonresponse, Zhang assumes that the probability to respond is logistic with a fixed composition effect $\xi_c$ and a random area effect $b_a$ as the explanatory variables (same probability for all the units in a given cell defined by area$\times$category). The model depends therefore on two sets of random effects, one set for the underlying complete data set with a vector of correlated multivariate normal composition effects in each area for the GSPREE model, and the other set for the response probabilities. Zhang (2009) estimates the small area compositions under the extended GSPREE using the EM algorithm, and estimates the prediction mean square errors under the model, accounting for the fixed and random effects estimation. The approach is applied to a real data set from Norway.

## 8. MODEL SELECTION AND CHECKING

Model selection and checking is one of the major problems in SAE because the models usually contain unobservable random effects, with limited or no information on their distribution. Notice that classical model selection criteria such as the AIC do not apply straightforwardly for mixed models because they use the likelihood, which requires specification of the distribution of the random effects, and because of difficulties in determining the effective sample size. In what follows I review several recent studies devoted to model selection and validation from both a frequentist and Bayesian perspective. These should be considered as supplements to 'ordinary' checking procedures based on graphical displays, significance testing, sensitivity of the computed predictors and their PMSEs to the choice of the likelihood and the prior distributions, and comparison of the model-dependent predictors with the corresponding model free direct estimators in sampled areas. Such model evaluation procedures can be found in almost every article on SAE, see, e.g., Mohadjer *et al*. (2007) and Nandram and Choi (2011) for recent diverse applications.

Vaida and Blanchard (2005) study the use of the AIC assuming the model (6.1) with $Var(u_i) = Q$, $Var(e_i) = \sigma^2 I_{n_i}$. The authors distinguish between inference on the marginal model with focus on the fixed effects, and inference on the model operating in the small areas with the associated vector random effects $u_i$. For the first case, the model can be written as a regression model with correlated residuals; $y_i = X_i\beta + v_i;\ v_i = Z_i u_i + e_i \sim N(0, Z_i Q Z_i' + \sigma^2 I_{n_i})$, in which case the classical (marginal) AIC, $mAIC = -2\log g(y \mid \hat{\psi}_{MLE}) + 2P$ applies, where $y$ is the vector of all the observations, $g(y \mid \hat{\psi}_{MLE})$ is the marginal likelihood evaluated at the MLE of $\psi$, the vector containing $\beta$, $\sigma^2$ and the unknown elements of $Q$, and $P = \dim(\psi)$. Gurka (2006) validates by simulations that one can use also in this case the mAIC with $\hat{\psi}_{REML}$, despite the use of different fixed effects design matrices under different models.

For the case where the focus is the model operating at the small areas, Vaida and Blanchard (2005) propose using a conditional AIC, which for a given model $g(y \mid \psi, u)$ is defined as,

$$(8.1) \qquad cAIC = -2\log g(y \mid \hat{\psi}_{MLE}, \hat{u}) + 2P^*;\ P^* = \frac{n(n-k-1)(\rho+1) + n(k+1)}{(n-k)(n-k-2)},$$

where $k$ is the number of covariates, $\hat{u} = E(u \mid \hat{\psi}_{MLE}, y)$ is the EBP of $u$ and $\rho = tr(H)$ with $H$ defining the matrix mapping the observed vector $y$ into the fitted vector $\hat{y} = X\hat{\beta} + Z\hat{u}$, such that $\hat{y} = Hy$. Notice that under this definition, the $u_i$ s are additional parameters. A conditional AIC for the case where $\psi$ is estimated by REML is also developed. The article contains theoretical results on properties of the cAIC and empirical results illustrating its good performance. The use of (8.1) is not restricted to mixed linear models with normal distributions of the error terms and it can be used to select the design matrices $X_i$ and $Z_i$.

Pan and Lin (2005) propose alternative goodness of fit test statistics for the GLMM based on estimated cumulative sums of residuals. Utilizing the notation for the model (6.1), the GLMM assumes the existence of a one to one link function $g(\cdot)$ satisfying $g[E(y_{ij} \mid u_i)] = \mathrm{x}_{ij}'\beta + \mathrm{z}_{ij}'u_i$, where $\mathrm{x}_{ij}$ and $\mathrm{z}_{ij}$ are the rows of the matrices $X_i$ and $Z_i$ corresponding to unit $(i, j) \in s_i$. The unconditional predictor of $y_{ij}$ is $m_{ij}(\psi) = E(y_{ij}) = E_{u_i}\{E[g^{-1}(x_{ij}'\beta + z_{ij}'u_i) \mid u_i]\}$, which is estimated by $m_{ij}(\hat{\psi})$. The estimated model residuals are therefore $e_{ij} = y_{ij} - m_{ij}(\hat{\psi})$ and they are computed by numerical integration. The authors consider two test statistics based on the distributions of aggregates of the residuals;

(8.2) $W(\mathrm{x}) = n^{-1/2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} I(\mathrm{x}_{ij} \le \mathrm{x})\, e_{ij};\;\; W_g(r) = n^{-1/2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} I(\hat{\mathrm{m}}_{ij} \le \mathrm{r})\, e_{ij}$,

where $I(\mathrm{x}_{ij} \le \mathrm{x}) = \prod_{l=1}^{k} I(x_{ijl} \le x_l)$. In particular, for testing the functional form of the *l*th covariate, one may consider the process $W_l(x) = n^{-1/2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} I(\mathrm{x}_{ijl} \le x)\, e_{ij}$, which is a special case of $W(\mathrm{x})$. The authors develop a simple approximation for the null distribution of $W_l(x)$ and use it for visual inspection by plotting the observed values against realizations from the null distributions for different values of $x$, and for a formal test defined by the supremum $S_l = \sup_x |W_l(x)|$. The statistic $S_l$ is used for testing the functional form of the deterministic part of the model. To test the appropriateness of the link function, the authors follow similar steps, using the statistics $W_g(r)$ for visual inspection and $S_g = \sup_r |W_g(r)|$ for formal testing. As discussed in the article, although different tests are proposed for different parts of the model, each test actually checks the entire model set up, including the assumptions regarding the random components.

The goodness-of-fit tests considered so far assume a given structure of the random effects, but are random effects actually needed in a SAE model applied to a given data set? Datta *et al*. (2011) show that if in fact the random effects are not needed and are removed from the model, this would improve the precision of point and interval estimators. The authors assume the availability of *k* covariates $\mathrm{x}_i = (x_{1i},...,x_{ki})$, $i = 1,...,m$ (viewed random for the theoretical developments) and weighted area-level means $\bar{y}_i = \sum_{j=1}^{n_i} w_{ij} y_{ij};\; \sum_{j=1}^{n_i} w_{ij} = 1$ of the outcome with known weights and known sums $W_{ir} = \sum_{j=1}^{n_i} w_{ij}^r,\; r = 2,...,q,\; q \le k$. The weights $w_{ij}$ are used for generating new area level means from bootstrap samples and the sums $W_{ir}$ are used for estimating model parameters by constructing appropriate estimating equations.

In order to test for the presence of random effects, the authors propose the test statistic

(8.3) $T = \sum_{i=1}^{m} [W_{i2}\lambda_2(\mathrm{x}_i, \hat{\psi})]^{-1} [\bar{y}_i - \lambda_1(\mathrm{x}_i, \hat{\psi})]^2$,

where $\lambda_l(\mathrm{x}_i, \hat{\psi}),\; l = 1, 2$ define the conditional mean and residual variance of $y \mid \mathrm{x}$ under the reduced model with no random effects and estimated (remaining) parameters $\hat{\psi}$. Critical values of the distribution of $T$ under the null hypothesis of no random effects are obtained by generating bootstrap samples with new outcomes from the conditional distribution of $y \mid \mathrm{x}; \hat{\psi}$ for given (original) covariates and weights, and computing the test statistic for each sample.

Empirical results indicate good powers of the proposed procedure and reduction in PMSE when the null hypothesis is not rejected. The procedure is applicable to very general models.

Jiang *et al*. (2008) propose a class of strategies for mixed model selection called *fence methods*, which apply to LMM and GLMM. The strategies involve a procedure to isolate a subgroup of correct models, and then select the optimal model from this subgroup according to some criterion. Let $Q_M = Q_M(y, \psi_M)$ define a measure of 'lack of fit' of a candidate model $M$ with parameters $\psi_M$, such that $E(Q_M)$ is minimized when $M$ is the true model. Examples of $Q_M$ are minus the loglikelihood or the residual sum of squares. Define $\hat{Q}_M = Q_M(y, \hat{\psi}_M)$ $= \inf_{\psi_M \in \Psi_M} Q_M(y, \psi_M)$ and let $\tilde{M} \in \mathrm{M}$ be such that $Q_{\tilde{M}} = \min_{M \in \mathrm{M}} \hat{Q}_M$ where M represents the set of candidate models. It is shown that under certain conditions $\tilde{M}$ is a correct model with probability tending to one. In practice there can be more than one correct model and a second step of the proposed procedure is to select an optimal model among models that are within a fence around $Q_{\tilde{M}}$. Examples of optimality criteria are minimal dimension or minimum PMSE. The fence is defined as $\hat{Q}_M \le \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M,\tilde{M}}$, where $\hat{\sigma}_{M,\tilde{M}}$ is an estimate of the standard deviation of $\hat{Q}_M - \hat{Q}_{\tilde{M}}$ and $c_n$ is a tuning coefficient that increases with the total sample size. Jiang *et al*. (2008) discuss alternative ways of computing $\hat{\sigma}_{M,\tilde{M}}$ and propose an adaptive procedure for choosing the tuning coefficient. The procedure consists of parametric bootstrapping new samples from the 'full' model, computing for every candidate model $M \in \mathrm{M}$ the empirical proportion $p^*(M, c_n)$ that it is selected by the fence method with a given $c_n$, computing $p^*(c_n) = \max_{M \in \mathrm{M}} p^*(M, c_n)$ and choosing $c_n$ that maximizes $p^*(c_n)$.

Jiang *et al*. (2008) apply the method for selecting the covariates in the area-level model (5.1) and the unit level model (5.3). Jiang *et al*. (2010) apply the method for selecting nonparametric P-spline models of the form (6.31). Selecting a model in this case requires selecting the degree of the spline $p$, the number of knots $K$ and a smoothing parameter $\lambda$ used for estimation of the model parameters.

So far I considered model selection and diagnostic procedures under the frequentist approach, but sound model checking is obviously required also under the Bayesian approach. Although this article is concerned with new developments, it is worth starting with a simulation procedure proposed by Dey *et al*. (1998) since it highlights a possible advantage of the Bayesian approach in this respect. Let $d$ define a discrepancy measure between the

assumed model and the data, such as the negative of the first stage likelihood of a hierarchical model. Denote by $y_{obs}$ the observed data and assume an *informative prior*. The procedure consists of generating a large number *R* of new data sets $y_{obs}^{(r)}, r = 1,...,R$ under the presumed model via Monte Carlo simulations and comparing the posterior distribution of $d \mid y_{obs}$ with the distributions of $d \mid y_{obs}^{(r)}$. Specifically, for each posterior distribution $f(d \mid y_{obs}^{(r)})$ compute the vector of quantiles $q^{(r)} = q_{\alpha_1}^{(r)},...,q_{\alpha_Q}^{(r)}$ (say $\alpha_1 = 0.025,...,\alpha_Q = 0.975$), compute $\overline{q} = \sum_{r=1}^{R} q^{(r)} / R$ and the Euclidean distances between $q^{(r)}$ and $\overline{q}$, and check whether the distance of the quantiles of the distribution of $d \mid y_{obs}$ from $\overline{q}$ is smaller or larger than, say, the 95$^{th}$ percentile of the $R$ distances.

Remark 9. The procedure is computationally intensive and it requires informative priors to allow generating new data sets, but it is very flexible in terms of the models tested and the discrepancy measure(s) used. A frequentist analogue via parametric bootstrap would require that the distribution of $d$ does not depend on the model parameters, or that the sample sizes are sufficiently large to permit ignoring parameter estimation.

Bayarri and Castellanos (2007) investigate Bayesian methods for *objective* model checking, which requires *noninformative priors* for the parameters $\psi$. The authors assume a given diagnostic statistic *T* (not a function of $\psi$) and consider two "surprise measures" of conflict between the observed data and the presumed model; the p-value $\Pr^{h(\cdot)}[T(y) \geq t(y_{obs})]$, and the relative predictive surprise $RPS = h[t(y_{obs})] / \sup_t[h(t)]$, where $h(t)$ is some specified distribution. Denote by $\theta$ the small area parameters. Writing $f(y) = \int f(y \mid \theta) g(\theta) d\theta$, it is clear that defining $h(t)$ requires integrating $\theta$ out of $f(y \mid \theta)$ with respect to some distribution for $\theta$. The prior $g(\theta)$ cannot be used since it is also improper and the authors consider three alternative solutions: 1- set the model hyper-parameters $\psi$ at their estimated value and integrate with respect to $g(\theta \mid \hat{\psi})$. This is basically an application of empirical Bayes and $h^{EB}(t) = \int f(t \mid \theta) g(\theta \mid \hat{\psi}) d\theta$. 2- integrate $\theta$ out by use of the posterior distribution $g(\theta \mid y_{obs})$. 3- Noticing that under the above two solutions the data are used both for obtaining a proper distribution for $\theta$ and for computing the statistic $t(y_{obs})$, the third solution removes the information in $t(y_{obs})$ from $y_{obs}$ by using the conditional likelihood $f(y_{obs} \mid t_{obs}, \theta)$. The resulting posterior distribution for $\theta$ is then used to obtain the

distribution $h(t)$, similarly to the previous cases. The specified distribution $h(t)$ under all the three cases nay not have a closed form, in which case it is approximated by MCMC simulations. See the article for details and for illustrations of the approach showing in general the best performance under the third solution.

Yan and Sedransk (2007) consider a specific model inadequacy namely, fitting models that do not account for all the hierarchical structure present, and like the last article restrict to noninformative priors. The authors consider two testing procedures, both based on the predictive posterior distribution $f(\tilde{y} \mid y_{obs}) = \int f(\tilde{y} \mid \psi) p(\psi \mid y_{obs}) d\psi$, where $\tilde{y}$ and $y_{obs}$ are assumed independent given $\psi$. The first procedure uses the posterior predictive p-values, $p_{ij} = \Pr(\tilde{y}_{ij} \le y_{ij} \mid y_{obs})$. The second procedure uses the p-values of a diagnostic statistic $t(\cdot)$ or a discrepancy measure $d(\cdot)$ (see above), for example, the p-values $\Pr[t(\tilde{y}) \ge t(y_{obs}) \mid y_{obs}]$. The authors analyse the simple case of a balanced sample where the fitted model is $y_{ij} \mid \mu, \phi \overset{iid}{\sim} N(\mu, \phi),\ i = 1,...,m,\ j = 1,...n_0$. It is shown that if the model is correct, then as $N = n_0 m \to \infty$ the distributions of $y_{obs}$ and $\tilde{y} \mid y_{obs}$ are the same, and the p-values $p_{ij}$ are distributed uniformly, as revealed in a Q-Q plot. On the other hand, if the true model is the two-level model $y_{ij} \mid \theta_i, \phi_0 \overset{iid}{\sim} N(\theta_i, \phi_0),\ \theta_i \mid \mu_0, A_0 \overset{iid}{\sim} N(\mu_0, A_0)$, then as $N \to \infty$ the mean and variance of the two models still agree but not the covariances, so that it is the ensemble of the $p_{ij}$'s or their Q-Q plot against the uniform distribution, but not individual p-values that permits distinguishing the two models. This, however, is only effective if the intra-cluster correlation is sufficiently high and the number of areas sufficiently small. Similar conclusions hold when comparing a two-stage hierarchical model with a three-stage model, and when applying the second testing procedure with the classical ANOVA $F$ test statistic as the diagnostic statistic, that is, when computing $\Pr[F(\tilde{y}) \ge F(y_{obs}) \mid y_{obs}]$.

Yan and Sedransk (2010) consider a third procedure for detecting a missing hierarchical structure that uses Q-Q plots of the predictive standardized residuals $r_{ij} = \dfrac{y_{ij} - E(\tilde{y}_{ij} \mid y_{obs})}{[Var(\tilde{y}_{ij} \mid y_{obs})]^{1/2}}$ against the standard normal distribution. The conditions under which the procedure performs well in detecting a misspecified hierarchy are the same as above.

Finally, I like to mention two articles that in a certain way bridge between the frequentist and Bayesian approaches for model selection. The idea here is to set up a noninformative

prior under the Bayesian approach so that the resulting posterior small area predictors have acceptable properties under the frequentist approach. This provides frequentist validation to the Bayesian methodology and the analyst may then take advantage of the flexibility of Bayesian inference by drawing observations from the posterior distribution of the area means. Both articles consider the area-level model (5.1) but the idea applies to other models.

Datta *et al*. (2005) assume a flat prior for $\beta$ and seek a prior $p(\sigma_u^2)$ satisfying $E(V_{iHB}) = PMSE[\hat{\theta}_i(\hat{\sigma}^2_{u,RE})] + o(m^{-1})$, where $V_{iHB} = Var(\theta_i \mid y_{obs})$ is the posterior variance of $\theta_i$ and $PMSE[\hat{\theta}_i(\hat{\sigma}^2_{u,RE})]$ is the frequentist PMSE of the EBLUP (or EB) when estimating $\sigma_u^2$ by REML. The expectation and PMSE are computed under the joint distribution of $\theta$ and $y$ under the model. The unique prior satisfying this requirement is shown to be,

(8.4) $$p_i(\sigma_u^2) \propto (\sigma_{Di}^2 + \sigma_u^2)^2 / \sum_{i=1}^m (\sigma_{Di}^2 + \sigma_u^2)^2 .$$

The prior is area specific in the sense that different priors are required for different areas.

Ganesh and Lahiri (2008) extend the results of Datta *et al*. (2005) to a weighted combination of the PMSE, thus obtaining a single prior for all the areas. The authors seek a prior which for a given set of weights $\{\omega_i\}$ satisfies,

(8.5) $$\sum_{i=1}^m \omega_i \{E(V_{iHB}) - PMSE[\hat{\theta}_i(\hat{\sigma}^2_{u,RE})]\} = o(1/m) .$$

The prior $p(\sigma_u^2)$ satisfying (8.5) is shown to be,

(8.6) $$P(\sigma_u^2) \propto \sum_{i=1}^m 1/(\sigma_{Di}^2 + \sigma_u^2)^2 / \sum_{i=1}^m \omega_i [\sigma_{Di}^2 / (\sigma_{Di}^2 + \sigma_u^2)]^2 .$$

By appropriate choice of the weights $\{\omega_i\}$, the prior (8.6) contains as special cases the flat prior $p(\sigma_u^2) = U(0, \infty)$, the prior developed by Datta *et al*. (2005) for a given area and the average moment matching prior (obtained by setting $\omega_i \equiv 1$).

## 9. CONCLUDING REMARKS

In this article I reviewed many new important developments in design- and model-based SAE. These developments give analysts much richer and versatile tools for their applications. Which approach should one follow in practice? Model-based predictors are generally more accurate and as discussed in Section 4.3, the models permit predictions for nonsampled areas for which no design-based theory exists. With everything else that can be done under a model, much of which reviewed in Sections 6-8, it seems to me that the choice between the two approaches is clear-cut, unless the sample sizes in all the areas are sufficiently large, although even in this case models have much more to offer like, for example, in the case of

measurement errors or informative nonresponse. This is not to say that design-based estimators have no role in model-based prediction. To begin with, the design-based estimators are often the input data for the model, as under the area-level model. Design-based estimators can be used for assessing the model-based predictors or for calibrating them via benchmarking, and the sampling weights play an important role when accounting for informative sampling within the sampled areas.

Next is the question of whether to follow the Bayesian approach (BA) or the frequentist approach (FA). I have to admit that before starting this extensive review I was very much in favor of FA, but the BA has some clear advantages. This is because one can generate as many observations as desired from the posterior distributions of the area parameters and hence it is much more flexible in the kind of models and inference possibilities that it can handle, for example, when the linking model does not match the conditional sampling model (Remark 2). Note also that the computation of PMSE (Bayes risk) or credibility intervals under BA does not rely on asymptotic results. A common criticism of BA is that it requires specification of prior distributions but as emphasized in Section 8, Bayesian models with proper, or improper priors can be tested in a variety of ways. Another criticism is that the application of BA is often very computation intensive and requires expert knowledge and computing skills even with modern available software. While this criticism may be correct (notably in my experience), the use of FA methods when fitting the GLMM is also very computation intensive and requires similar skills. Saying all this, it is quite obvious to me that the use of FA will continue to be dominant for many years to come because except for few exceptions, official statistical bureaus are very reluctant to use Bayesian methods.

Where do we go from here? Research on SAE continues all over the world, both in terms of new theories and in applications to new intriguing problems and I hope that this review will contribute to this research. The new developments that I have reviewed are generally either under BA or FA, and one possible direction that I hope to see is to incorporate the new developments under one approach into the other. For example, use the EL approach under FA, use spline regressions under BA, account for informative nonresponse in FA, or produce poverty mapping with BA. Some of these extensions will be simple; other may require more extensive research and some may not be feasible, but this will make it easier for analysts to choose between the two approaches.

REFERENCES


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

BAYARRI, M. J. and CASTELLANOS, M.E. (2007). Bayesian checking of the second levels of Hierarchical Models. *Statistical Science* **22** 322-343.

BELL, W. R. and HUANG, E. T. (2006). Using the *t*-distribution to deal with outliers in small area estimation. Proceedings of Statistics Canada Symposium on Methodological Issues in Measuring Population Health.

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

CHAMBERS, R., CHANDRA, H. and TZAVIDIS, N. (2011). On bias-robust mean squared error estimation for pseudo-linear small are estimators. *Survey Methodology* **37** 153-170.

CHANDRA, H. and CHAMBERS, R. (2009). Multipurpose small area estimation. *Journal of Official Statistics* **25** 379–395.

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

CHAUDHURI, S. and GHOSH, M. (2011). Empirical likelihood for small area estimation. *Biometrika* **98** 473-480.

CHEN, S. and LAHIRI, P. (2002). On mean squared prediction error estimation in small area estimation problems. In Proceedings of the Survey Research Methods Section, American Statistical Association. pp. 473-477.

DAS, K., JIANG, J. and RAO, J.N.K. (2004). Mean squared error of empirical predictor. *Annals of Statistics*, **32**, 818-840.

DATTA, G. S. (2009). Model-Based Approach to Small Area Estimation. In: Handbook of Statistics 29B; *Sample Surveys*: *Inference and Analysis*. Eds. D. Pfeffermann and C.R. Rao. North Holland. pp. 251-288.

DATTA, G. S. and LAHIRI, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica* **10** 613–627.

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

DATTA, G. S., RAO, J. N. K. and 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**, 2952-2963.

DATTA, G. S., GHOSH, M., STEORTS, R. and MAPLES, J. (2011). Bayesian Benchmarking with Applications to Small Area Estimation. *Test* **20** 574-588.

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

DEY, D.K., GELFAND, A.E., SWARTZ, T.B. and VLACHOS, A.K (1998). A simulation-intensive approach for checking hierarchical models. *Test* **7** 325-346.

ESTEVAO, V. M. and SÄRNDAL, C. E. (2004). Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *Journal of Official Statistics* **20** 645–669.

ESTEVAO, V. M. and SÄRNDAL, C. E. (2004). Survey estimates by calibration on complex auxiliary information. *International Statistical Review* **74** 127-147.

FALORSI, P. D. and RIGHI, P. (2008). A balanced sampling approach for multi-way stratification designs for small area estimation. *Survey Methodology* **34** 223-234.

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

GANESH, N. and LAHIRI, P. (2008). A new class of average moment matching prior. *Biometrika* **95** 514-520.

GHOSH, M., NATARAJAN, K., STROUD, T.W. and CARLIN, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association* **93**, 273-282.

GHOSH, M., SINHA, K. and KIM, D. (2006). Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error models. *Scandinavian Journal of Statistics* **33**, 591-608.

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

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

GURKA, M.J. (2006). Selecting the best linear mixed model under REML. *The American Statistician* **60** 19-26.

HALL, P. and MAITI, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society Series B* **68** 221- 238.

HUANG, E.T. and BELL, W.R. (2006). Using the *t*-distribution in small area estimation: an application to SAIPE state poverty models. In: Proceedings of the Survey Research Methods Section, American Statistical Association. pp. 3142-3149.

JIANG, J., LAHIRI, P. S. and WAN, S. M. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics* **30** 1782–1810.

JIANG, J. and LAHIRI, P. S. (2006a). Estimation of finite population domain means: a model-assisted empirical best prediction approach. *Journal of the American Statistical Association* **101** 301–311.

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

JIANG, J., RAO, J. S., GU, Z. and NGUYEN, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics* **36** 1669-1692.

JIANG, J., NGUYEN, T. and RAO, J. S. (2010). Fence method for nonparametric small area estimation. *Survey Methodology* **36** 3-11.

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

KOTT, P. S. (2009). Calibration Weighting: Combining Probability Samples and Linear Prediction Models. In: Handbook of Statistics 29B; *Sample Surveys*: *Inference and Analysis*. Eds. D. Pfeffermann and C.R. Rao. North Holland. pp. 55-82.

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

LEHTONEN, R., SÄRNDAL, C. E. and VEIJANEN, A. (2005). Does the model matter? comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition* **7** 649–673.

LEHTONEN, R. and VEIJANEN, A. (2009). Design-based Methods of Estimation for Domains and Small Areas. In: Handbook of Statistics 29B; *Sample Surveys*: *Inference and Analysis*. Eds. D. Pfeffermann and C.R. Rao. North Holland. pp. 219-249.

LOHR, S. L. and RAO, J. N. K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika* **96** 457-468.

MACGIBBON, B. and TOMBERLIN, T. J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology* **15** 237-252.

MALEC, D., DAVIS, W. W. and CAO, X. (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine* **18** 3189–3200.

MALINOVSKY, Y. and RINOTT, Y. (2010). Prediction of ordered random effects in a simple small area model. *Statistica Sinica* **20** 697-714.

MOHADJER, L., RAO, J.N.K., LIU, B., KRENZKE, T. and VAN DE KERCKHOVE, W. (2007). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. In: Proceedings of the Survey Research Methods Section, American Statistical Association. pp. 3203-3210.

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

NANDRAM, B. and SAYIT, H. (2011). A Bayesian analysis of small area probabilities under a constraint. Survey Methodology **37** 137-152.

OPSOMER, J. D., CLAESKENS, G., RANALLI, M. G., KAUEMANN, G. and BREIDT, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society*, series B **70** 265-286.

PFEFFERMANN, D. (2002). Small Area Estimation- New Developments and Directions. I*nternational Statistical Review* **7**0 125-143.

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

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

PFEFFERMANN, D., TERRYN B. and 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* **34** 235-249.

PFEFFERMANN, D. and CORREA, S. (2012). Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. *Biometrika*, advanced published access at doi (10.1093.biomet/ass010).

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

RAO, J. N. K. (2003). *Small area estimation*. John Wiley & Sons, New York.

RAO, J. N. K., SINHA, S. K. and ROKNOSSADATI M. (2009). Robust small area estimation using penalized spline mixed models. In: Proceedings of the Survey Research Methods Section, American Statistical Association. pp. 145-153.

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

TORABI, M. and Rao, J. N. K. (2008). Small area estimation under a two-level model. *Survey Methodology* **34** 11–17.

TORABI, M., DATTA, G. S. and 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., MARCHETTI, S. and CHAMBERS, R. (2010). Robust estimation of small-areas means and quantiles. Australian & New Zealand Journal of Statistics 52 167-186.

UGARTE, M.D., MILITINO, A.F. and GOICOA, T. (2009). Benchmarked estimates in small areas using linear mixed models with restrictions. *Test* **18** 342-364.

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

WRIGHT, D.L., STERN, H.S. and CRESSIE, N. (2003). Loss function for estimation of extreme with an application to disease mapping. *The Canadian Journal of Statistics* **31** 251-266.

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

YAN, G. and SEDRANSK, J. (2007). Bayesian diagnostic techniques for detecting hierarchical structure. *Bayesian Analysis* **2** 735-740.

YAN, G. and SEDRANSK, J. (2010). A note on Bayesian residuals as a hierarchical model diagnostic technique. *Statistical Papers* **51** 1-10.

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

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

ZHANG L.C. and CHAMBERS, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society series B* **66** 479-496.

ZHANG L.C. (2009). Estimates for small area compositions subjected to informative missing data. *Survey Methodology* **35** 191-201.