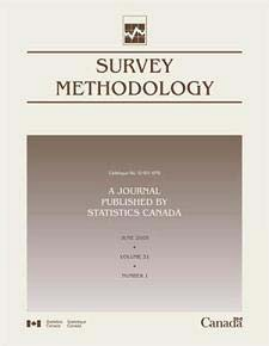
Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Some recent advances in model-based small area estimation



by J.N.K.Rao

December 1999

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 175-186
Statistics Canada, Catalogue No. 12-001

Some recent advances in model-based small area estimation

J.N.K. Rao [1]

Abstract

Small area estimation has received a lot of attention in recent years due to growing demand for reliable small area estimators. Traditional area-specific direct estimators do not provide adequate precision because sample sizes in small areas are seldom large enough. This makes it necessary to employ indirect estimators that borrow strength from related areas; in particular, model-based indirect estimators. Ghosh and Rao (1994) provided a comprehensive review and appraisal of methods for small area estimation, covering the literature to 1992-1993. This paper supplements Ghosh and Rao (1994) by covering the literature over the past five years or so on model-based estimation. In particular, we cover several small area models and empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes (HB) methods applied to these models. We also present several recent applications of small area estimation.

Key Words: Empirical Bayes; Hierarchical Bayes; Small area models.

1. Introduction

Sample surveys are used to provide estimates not only for the total population but also for a variety of subpopulations (domains). "Direct" estimators, based only on the domain-specific sample data, are typically used to estimate parameters for large domains. But sample sizes in small domains, particularly small geographical areas, are rarely large enough to provide direct estimates for specific small domains. For example, the U.S. Third National Health and Nutrition Examination Survey was designed to provide direct estimates with acceptable precision for domains classified by race, ethnicity and age. But, to have a large enough sample to support reliable direct estimates for, say, all states is seldom possible, and for all subareas like counties is almost never possible. In this example, states/counties may be regarded as "small areas" because the area-specific sample sizes are small (or even zero). In making estimates for such small areas it is necessary to "borrow strength" from related areas to form "indirect" estimators that increase the effective sample size and thus increase the precision. Such indirect estimators are based on either implicit or explicit models that provide a link to related small areas through supplementary data such as recent census counts and current administrative records. Indirect estimators based on implicit models include synthetic and composite estimators, while those based on explicit models incorporating area-specific effects include empirical Bayes (EB), empirical best linear unbiased prediction (EBLUP) and hierarchical Bayes (HB) estimators.

Ghosh and Rao (1994) presented a comprehensive overview and appraisal of methods for small area estimation, covering the literature to 1992-1993. We refer the reader to Schaible (1996) for an excellent account of the use of indirect estimators in U.S. Federal Programs.

Ghosh and Rao (1994) provided a list of symposia and workshops on small area estimation that have been organized in recent years. We update that list by the following: (i) Conference on Small Area Estimation, U.S. Bureau of the Census, Washington, D.C., March 26-27, 1998; and (ii) International Satellite Conference on Small Area Estimation, Riga, Latvia, August 20-21, 1999. Short courses have also been organized: (i) "Small Area Estimation" by J.N.K. Rao, W.A. Fuller, G. Kalton and W.L. Schaible, organized by the Joint Program in Survey Methodology and the Washington Statistical Society, Washington, D.C., May 22-23, 1995; and (ii) "Introduction to Small Area Estimation" by J.N.K. Rao, organized by the International Association of Survey Statisticians, Riga, Latvia, August 19, 1999. In addition, numerous invited and contributed sessions on small area estimation have been organized at recent professional statistical meetings, including the American Statistical Association Annual Meetings and the International Statistical Institute bi-annual sessions.

Singh, Gambino and Mantel (1994) discussed survey design issues that have an impact on small area statistics. In particular, they presented an excellent illustration of compromise sample size allocations to satisfy reliability requirements at the provincial level as well as sub provincial level. For the Canadian Labour Force Survey with a monthly sample of 59,000 households, optimizing at the provincial level yields a coefficient of variation (CV) for "unemployed" as high as 17.7% for some Unemployment Insurance (UI) regions. On the other hand, a two-step allocation with 42,000 households allocated at the first step to get reliable provincial estimates and the remaining 17,000 households allocated in the second step to produce best possible UI region estimates reduces the worst case of 17.7% CV for UI regions to 9.4% at the expense of a small increase in CV at the provincial and national levels: CV for Ontario increases from 2.8% to 3.4%

1. J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.

and for Canada from 1.36% to 1.51%. Preventive measures, such as compromise sample allocations, should be taken at the design stage, whenever possible, to ensure precision for domains like the UI region. But even after taking such measures sample sizes may not be large enough for direct estimates to provide adequate precision for all small areas of interest. As noted before, sometimes the survey is deliberately designed to oversample specific areas (domains) at the expense of small samples or even no samples in other areas of interest.

This paper supplements Ghosh and Rao (1994) by covering the literature over the past five years or so on model-based small area estimation; in particular, on empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB and hierarchical Bayes (HB) methods and their applications.

2. Small area models

It is now generally accepted that when indirect estimates are to be used they should be based on explicit models that relate the small areas of interest through supplementary data such as last census data and current administrative data. An advantage of the model approach is that it permits validation of models from the sample data. Interesting work on traditional indirect estimates (synthetic, sample-size dependent *etc*.), however, is also reported in the recent literature (see *e.g*., Falorsi, Falorsi and Russo 1994; Chaudhuri and Adhikary 1995; Schaible 1996; Marker 1999).

Small area models may be broadly classified into two types: area level and unit level.

2.1 Area level models

Area-specific auxiliary data, $\mathbf{x}_i$, are assumed to be available for the sampled areas $i(=1, ..., m)$ as well as the nonsampled areas. A basic area level model assumes that the population small area mean $\overline{Y}_i$ or some suitable function $\theta_i = g(\overline{Y}_i)$, such as $\theta_i = \log(\overline{Y}_i)$, is related to $\mathbf{x}_i$ through a linear model with random area effects v_i:

$$\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i, \qquad i = 1, ..., m \tag{2.1}$$

where $\boldsymbol{\beta}$ is the p-vector of regression parameters and the v_i's are uncorrelated with mean zero and variance σ_v^2. Normality of the v_i is also often assumed. The model (2.1) also holds for the non sampled areas. It is also possible to partition the areas into groups and assume separate models of the form (2.1) across groups.

We assume that direct estimators $\hat{\overline{Y}}_i$ of $\overline{Y}_i$ are available whenever the area sample size $n_i \geq 1$. It is also customary to assume that

$$\hat{\theta}_i = \theta_i + e_i \tag{2.2}$$

where $\hat{\theta}_i = g(\hat{\overline{Y}}_i)$ and the sampling errors e_i are independent $N(0, \psi_i)$ with known ψ_i. Combining this sampling model with the "linking" model (2.1), we get the well-known area level linear mixed model of Fay and Herriot (1979):

$$\hat{\theta}_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i + e_i. \tag{2.3}$$

Note that (2.3) involves both design-based random variables e_i and model-based random variables v_i. In practice, sampling variances ψ_i are seldom known, but smoothing of estimated variances $\hat{\psi}_i$ is often done to get stable estimates ψ_i^* which are then treated as the true ψ_i. Other methods of handling unknown ψ_i are mentioned in section 3.4. An advantage of the area-level model (2.3) is that the survey weights are accounted for through the direct estimators $\hat{\theta}_i$.

The assumption $E(e_i | \theta_i) = 0$ in the sampling model (2.2) may not be valid if the sample size n_i is small and θ_i is a nonlinear function of the total Y_i, even if the direct estimator $\hat{Y}_i$ is design-unbiased, *i.e*., $E(\hat{Y}_i | Y_i) = Y_i$. A more realistic sampling model is given by

$$\hat{Y}_i = Y_i + e_i^* \tag{2.4}$$

with $E(e_i^* | Y_i) = 0$, *i.e*., $\hat{Y}_i$ is design-unbiased for the total Y_i. In this case, however, we cannot combine (2.4) with the linking model to produce a linear mixed model. As a result, standard results in linear model theory do not apply, unlike in the case of (2.3). Alternative methods to handle this case are needed (see section 4.1).

The basic area level model has been extended to handle correlated sampling errors, spatial dependence of random small area effects, vectors of parameters $\boldsymbol{\theta}_i$ (multivariate case), time series and cross-sectional data and others (see Ghosh and Rao 1994). We discuss some of the recent models for combining cross-sectional and time series data. Suppose θ_{it} denotes a parameter of interest for small area i at time t and $\hat{\theta}_{it}$ is a direct estimator of θ_{it}. Ghosh, Nangia and Kim (1996) assumed the sampling model $\hat{\theta}_{it} | \theta_{it} \overset{\text{ind}}{\sim} N(\theta_{it}, \psi_{it})$ with known sampling variances ψ_{it}, and the linking model

$$\theta_{it} | \mathbf{u}_t \sim N(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{u}_t, \sigma_t^2) \tag{2.5}$$

and

$$\mathbf{u}_t | \mathbf{u}_{t-1} \sim N(\mathbf{u}_{t-1}, \mathbf{W}) \tag{2.6}$$

with known auxiliary variables $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$; they have actually studied the multivariate case $\boldsymbol{\theta}_{it}$. Note that (2.6) is the well-known random walk model. The above model has the following limitations: (i) Independence of $\hat{\theta}_{it}$'s over t for each i may not be realistic because estimates are typically correlated over time. (ii) The linking model (2.5) does not include area-specific random effects. As a result, it is likely to produce oversmooth estimates. Rao and Yu (1992, 1994) proposed more realistic sampling and linking models. They assumed the sampling model

$$\hat{\boldsymbol{\theta}}_i | \boldsymbol{\theta}_i \overset{\text{ind}}{\sim} N(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i) \tag{2.7}$$

with known sampling covariance matrix ψ_i, and the linking model

$$\theta_{it} = \mathbf{x}_{it}^T\boldsymbol{\beta} + v_i + u_{it} \quad (2.8)$$

with $v_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2)$ and independent of u_{it}'s which are assumed to follow an AR(1) model:

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, \quad |\rho| < 1 \quad (2.9)$$

with $\varepsilon_{it} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, where $\hat{\boldsymbol{\theta}}_i = (\hat{\theta}_{i1}, ..., \hat{\theta}_{iT})'$ and $\boldsymbol{\theta}_i = (\theta_1, ..., \theta_{iT})'$. Models of the form (2.7)-(2.9) have been extensively studied in the econometric literature, ignoring sampling errors, *i.e.*, treating $\hat{\theta}_{it}$ as θ_{it}. The above sampling model permits correlations among sampling errors over time and the linking model (1.9) includes both area-specific random effects v_i and area by time specific random effects u_{it}. Datta, Lahiri and Lu (1994), following Rao and Yu (1992), used the same sampling model (2.7) but assumed the following linking model:

$$\theta_{it} | v_i, \mathbf{u}_t \sim N(\mathbf{x}_{it}^T\boldsymbol{\beta}_i + v_i + \mathbf{z}_{it}^T\mathbf{u}_t, \sigma_i^2) \quad (2.10)$$

where $\boldsymbol{\beta}_i$'s and σ_i^2's are random and $\mathbf{u}_t$ follows the random walk model (2.6). This model allows area-specific random effects v_i and random slopes $\boldsymbol{\beta}_i$, but does not contain area by time specific random effects u_{it}. Datta, Lahiri and Maiti (1999) used the Rao-Yu sampling and linking models (2.7) and (2.8) but replaced the AR(1) model (2.9) by a random walk model given by (2.9) with $\rho = 1$. Datta, Lahiri, Maiti and Lu (1999) considered a similar model but added extra terms to $\mathbf{x}_{it}^T\boldsymbol{\beta} + v_i$ to reflect seasonal variation in their application to estimating unemployment rates for the U.S states. Singh, Mantel and Thomas (1994) also used time series/cross-sectional models, but assumed that the sample errors are uncorrelated over time.

Area level models have also been used in the context of disease mapping or estimating regional mortality and disease rates, as noted by Ghosh and Rao (1994). A simple model assumes that the observed small area disease counts $y_i | \theta_i \overset{\text{ind}}{\sim}$ Poisson $P(n_i\theta_i)$ and $\theta_i \overset{\text{i.i.d.}}{\sim}$ gamma $G(a, b)$, where θ_i is the true incidence rate and n_i is the number exposed in area i. Maiti (1998) used $\beta_i = \log\theta_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ instead of $\theta_i \overset{\text{i.i.d.}}{\sim} G(a, b)$. He also considered a spatial dependence model for β_i's, using conditional autoregression (CAR) that relates each β_i to a set of neighbourhood areas of area i; see also Ghosh, Natarajan, Kim and Walker (1997). Lahiri and Maiti (1996) modelled age-group specific area disease counts y_{ij}, using Clayton and Kaldor's (1987) approach. They assumed that $y_{i.} = \sum_j y_{ij} | \theta_i \overset{\text{ind}}{\sim} P(e_i\theta_i)$ and $\theta_i \overset{\text{i.i.d.}}{\sim} G(a, b)$, where $e_i = \sum_j \psi_j n_{ij}$ is the expected number of deaths in area i, ψ_j is the j^{th} group effect assumed to be known and n_{ij} is the number exposed in the j^{th} age group and area i. Nandram, Sedransk and Rickle (1998) assumed that $y_{ij} | \theta_{ij} \overset{\text{ind}}{\sim} P(n_{ij}\theta_{ij})$ and $\log\theta_{ij} = \mathbf{x}_j'\boldsymbol{\beta} + v_i$ with $v_i \overset{\text{ind}}{\sim} N(0, \sigma^2)$, where θ_{ij} is the area/age-specific mortality rate and $\mathbf{x}_j$ is a vector of covariates for age group j. They also considered random slopes $\boldsymbol{\beta}_i$ in the linking model.

2.2 Unit level models

A basis unit level population model assumes that the unit y-values y_{ij}, associated with the units j in the areas i, are related to auxiliary variables $\mathbf{x}_{ij}$ through a one-way nested error regression model

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1, ..., N_i;\ i = 1, ..., m \quad (2.11)$$

where $v_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2)$ are independent of $e_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_e^2)$ and N_i is the number of population units in the i^{th} area. The parameters of interest are the totals Y_i or the means $\bar{Y}_i$.

The model (2.11) is appropriate for continuous variables y. To handle count or categorical (*e.g.*, binary) y-variables, generalized linear mixed models with random small area effects, v_i, are often used. Ghosh, Natarajan, Stroud and Carlin (1998) assumed models of the form: (i) Given θ_{ij}'s, the y_{ij}'s are independent and belong to the exponential family with canonical parameter θ_{ij}; (ii) Linking model $g(\theta_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i$ where $v_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2)$ and $g(\cdot)$ is a strictly increasing function. The linear mixed model (2.11) is a special case of this class with $g(a) = a$. The logistic function $g(a) = \log[a/(1-a)]$ is often used for binary y (see *e.g.*, Farrell, McGibbon and Tomberlin 1997) although probit functions can also be used and offer certain advantages for hierarchical Bayes (HB) inference (Das, Rao and You 1999).

The sample data $\{y_{ij}, \mathbf{x}_{ij}, j = 1, ..., n_i;\ i = 1, ..., m\}$ is assumed to obey the population model. This implies that the sample design is ignorable or selection bias is absent which is satisfied by any equal probability sampling method within areas. For more general designs, the sample indicator variable, a_{ij}, should be unrelated to y_{ij}, conditional on x_{ij}. Model-based estimators for unit level models do not depend on the survey weights, $\tilde{w}_{ij}$, so that design-consistency as n_i increases is forsaken except when the design is self-weighting, *i.e.*, $\tilde{w}_{ij} = \tilde{w}$, as in the case of equal probability sampling. The area level model (2.3) is free of these limitations but assumes that the sample variances ψ_i are known; if ψ_i's are assumed unknown the model becomes nonidentifiable or nearly nonidentifiable leading to highly unstable estimates of the parameters. The unit level model is free of the latter difficulty and survey weights can also be incorporated using model-assisted estimators; see the paragraph containing equation (3.8).

Various extensions of the basic unit level models have been studied over the past five years or so. Stukel and Rao (1999) studied two-way nested error regression models which are appropriate for two-stage sampling within small areas. Following Kleffe and Rao (1992), Arora and Lahiri (1997) studied unit level models of the form (2.11) with random error variances σ_i^2 such that $\sigma_i^{-2} \overset{\text{i.i.d.}}{\sim} G(a, b)$; Kleffe and Rao (1992) assumed the existence of only mean and variance of σ_i^2, without specifying a parametric distribution

on σ_i^2. Datta, Day and Basawa (1999) extended the unit level model (2.11) to the multivariate case y_{ij}, following Fuller and Harter (1987). This extension leads to a multivariate nested error regression model. Moura and Holt (1999) generalized (2.11) to allow some or all of the regression coefficients to be random and to depend on area level auxiliary variables, thus effectively integrating the use of unit level and area level covariates into a single model. You and Rao (1999a) also studied similar two-level models.

Malec, Davis and Cao (1996, 1999) and Malec, Sedransk, Moriarity and LeClere (1997) studied the binary case, using logistic linear mixed models with random slopes to link the small areas. Raghunathan (1993) specified only the first two moments of y_{ij}'s conditional on small area means θ_i's and the first moment of θ_i as $\tau_i = h(\mathbf{z}_i'\boldsymbol{\beta})$ for known inverse "link" function $h(\cdot)$ and the second moment of θ_i is allowed to depend on τ_i.

Many of the small area linear mixed models studied in the literature are special cases of the following general linear mixed model with a block diagonal covariance structure, sometimes called longitudinal mixed linear models (Prasad and Rao 1990; Datta and Lahiri 1997):

$$\mathbf{y}_i^* = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{v}_i + \mathbf{e}_i, \quad i = 1, ..., m \tag{2.12}$$

where $\mathbf{v}_i \overset{\text{ind}}{\sim} (\mathbf{0}, \mathbf{G}_i(\boldsymbol{\tau}))$ and independent of $\mathbf{e}_i \overset{\text{ind}}{\sim} (\mathbf{0}, \mathbf{R}_i(\boldsymbol{\tau}))$. For example, the basic area level (2.3) is of the form (2.12) with $\mathbf{y}_i^* \hat{\theta}_i, \mathbf{Z}_i = 1, \mathbf{G}_i(\boldsymbol{\tau}) = \sigma_v^2$ and $\mathbf{R}_i(\boldsymbol{\tau}) = \psi_i$. Das, Rao and You (1999) studied general mixed ANOVA models of the form

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{v}_1 + ... + \mathbf{Z}_q\mathbf{v}_q + \mathbf{e}_i \tag{2.13}$$

where $\mathbf{Z}_i$ consists of only 0's and 1's such that there is exactly one 1 in each row and at least one 1 in each column, $\mathbf{v}_i \overset{\text{ind}}{\sim} (\mathbf{0}, \sigma_i^2\mathbf{I})$ and independent of $\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I})$. This model relaxes the assumption of a block diagonal covariance structure.

Ghosh and Rao (1994) reviewed some work on model diagnostics for models involving random effects. Jiang, Lahiri and Wu (1998) developed a chi-squared test for checking the normality of the random effects v_i and the errors e_{ij} in the basic unit level sample model $y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{v}_i + e_{ij}$, $j = 1, ..., n_i$; $i = 1, ..., m$.

3. Model-based inference: Basic area-level model

EBLUP, EB and HB methods have played a prominent role for model-based small area estimation. EBLUP is applicable for linear mixed models whereas EB and HB are more generally valid. EBLUP point estimators do not require distributional assumptions, but normality of random effects is often assumed for estimating the mean squared error (MSE) of the estimators. Also, EBLUP and EB estimators are identical under normality and nearly equal to the HB estimator, but measures of variability of the estimators may be different. To illustrate the methods, we focus on the basic area level model (2.3), which is extensively used in practice. Various extensions of the basic area-level and unit level models are studied in section 4.

3.1 EBLUP method

Appealing to general results for linear mixed models, the BLUP estimator of θ_i under (2.3) is given by

$$\tilde{\theta}_i(\sigma_v^2) = \gamma_i\hat{\theta}_i + (1-\gamma_i)\mathbf{x}_i'\tilde{\boldsymbol{\beta}}(\sigma_v^2) \tag{3.1}$$

where $\gamma_i = \sigma_v^2/(\sigma_v^2 + \psi_i)$ and $\tilde{\boldsymbol{\beta}}(\sigma_v^2)$ is the weighted least squares (WLS) estimator of $\boldsymbol{\beta}$ with weights $(\sigma_v^2 + \psi_i)^{-1}$. It follows from (3.1) that the BLUP estimator is a weighted combination of the direct estimator $\hat{\theta}_i$ and the regression synthetic estimator $\mathbf{x}_i'\tilde{\boldsymbol{\beta}}(\sigma_v^2)$. The result (3.1) does not require the normality of v_i and e_i. Since σ_v^2 is unknown, we replace it by a suitable estimator $\hat{\sigma}_v^2$ to obtain a two-step or EBLUP estimator $\tilde{\theta}_i = \tilde{\theta}_i(\hat{\sigma}_v^2)$. The estimator of total Y_i is taken as $g^{-1}(\tilde{\theta}_i) = h(\tilde{\theta}_i)$. One could use either the method of fitting constants (not requiring normality) or the restricted maximum likelihood (REML) method under normality to estimate σ_v^2. Jiang (1996) showed that REML estimators of variance components in linear mixed models remain consistent under deviations from normality. Therefore, $\tilde{\theta}_i$ with REML estimator of σ_v^2 is also asymptotically valid under nonnormality.

As noted in section 2.1, EBLUP estimation is not applicable if the sampling model (2.2) is changed to the more realistic model (2.4).

A measure of variability associated with EBLUP estimator is given by its MSE, but no closed form for MSE exists except in some special cases. As a result, considerable attention has been given in recent years to obtain accurate approximations to the MSE of EBLUP estimators. An accurate approximation to $\text{MSE}(\tilde{\theta}_i) = E(\tilde{\theta}_i \sim \theta_i)^2$, for large m, under normality is given by

$$\text{MSE}(\tilde{\theta}_i) \approx g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2) \tag{3.2}$$

where

$$g_{1i}(\sigma_v^2) = \gamma_i\psi_i, \tag{3.3}$$

$$g_{2i}(\sigma_v^2) = (1-\gamma_i)^2\mathbf{x}_i'\left[\sum_i \mathbf{x}_i\mathbf{x}_i'/(\sigma_v^2 + \psi_i)\right]^{-1}\mathbf{x}_i, \tag{3.4}$$

$$g_{3i}(\sigma_v^2) = [\psi_i^2/(\sigma_v^2 + \psi_i)^4]E(\hat{\theta}_i - \mathbf{x}_i'\boldsymbol{\beta})^2\overline{V}(\hat{\sigma}_v^2), \tag{3.5}$$

$$= [\psi_i^2/(\sigma_v^2 + \psi_i)^2]\,\overline{V}(\hat{\sigma}_v^2) \tag{3.6}$$

and $\overline{V}(\hat{\sigma}_v^2)$ is the asymptotic variance of $\hat{\sigma}_v^2$ (Prasad and Rao 1990). The leading term $g_{1i}(\sigma_v^2) = \gamma_i\psi_i$ is of order $O(1)$

whereas $g_{2i}(\sigma_v^2)$, due to estimating $\boldsymbol{\beta}$, and $g_{3i}(\sigma_v^2)$, due to estimating σ_v^2, are both of order $O(m^{-1})$, for large m. Note that the leading term shows that $\text{MSE}(\tilde{\theta}_i)$ can be substantially smaller than $\text{MSE}(\hat{\theta}_i)$ under the model (2.3) when γ_i is small or the model variance σ_v^2 is small relative to the sampling variance ψ_i. The success of small area estimation, therefore, largely depends on getting good auxiliary information $\{\mathbf{x}_i\}$ that leads to a small model variance relative of ψ_i. Of course, one should also make a thorough validation of the assumed model.

An estimator of $\text{MSE}(\tilde{\theta}_i)$, correct to the same order of approximation as (3.2), is given by

$$\text{mse}(\tilde{\theta}_i) \approx g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2), \qquad (3.7)$$

i.e., the bias of (3.7) is of lower order than m^{-1} for large m. The approximation (3.7) is valid for both the method of fitting constants estimator and the REML estimator, but not for the ML estimator of σ_v^2 (Datta and Lahiri 1997; Prasad and Rao 1990). Using the fitting of constants estimator, Lahiri and Rao (1995) showed that (3.7) is robust to non-normality of the small area effects v_i in the sense that approximate unbiasedness remains valid. Note that the normality of sampling errors e_i is still assumed but it is less restrictive due to the central limit theorem's effect on the direct estimators $\hat{\theta}_i$.

A criticism of the MSE estimator (3.7) is that it is not area-specific in the sense that it does not depend on $\hat{\theta}_i$ although $\mathbf{x}_i$ involved through (3.4). But it is easy to find other choices using the form (3.5) for $g_{3i}(\sigma_v^2)$. For example, we can use

$$\begin{aligned}\text{mse}_1(\tilde{\theta}_i) &= g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + g_{3i}(\hat{\sigma}_v^2) \\ &= [\psi_i^2/(\hat{\sigma}_v^2 + \psi_i)^4]\,(\hat{\theta}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2 h_i(\hat{\sigma}_v^2), \qquad (3.8)\end{aligned}$$

where $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_v^2)$ and $h_i(\sigma_v^2) = \bar{V}(\hat{\sigma}_v^2) = 2m^{-2}\sum_i (\sigma_v^2 + \psi_i)^2$ for the fitting of constants estimator $\hat{\sigma}_v^2$ (Rao 1998). The last term of (3.8) is less stable than $g_{3i}(\hat{\sigma}_v^2)$ but it is of lower order than the leading term $g_{1i}(\hat{\sigma}_v^2)$.

Rivest and Belmonte (1999) obtained an unbiased estimator of the conditional MSE of the EBLUP estimator $\tilde{\theta}_i = \tilde{\theta}_i(\hat{\sigma}_v^2)$ for the basic area level model, assuming only the sampling model, *i.e.*, conditionally given θ_i's. Hwang and Rao (1987) obtained similar results and showed empirically that the model-based estimator of MSE, (3.7), is much more stable than the unbiased estimator and that it tracks the conditional MSE quite well even under moderate violations of the assumed linking model (2.1). Only in extreme cases, such as large outliers θ_i, the model-based estimator might perform poorly compared to the unbiased estimator.

3.2 EB method

In the EB approach to the basic area level model, given by (2.1) and (2.2), the conditional distribution of θ_i given $\hat{\theta}_i$ and model parameters $\boldsymbol{\beta}$ and σ_v^2, denoted $f(\theta_i \mid \hat{\theta}_i, \boldsymbol{\beta}, \sigma_v^2)$, is first obtained. The model parameters are estimated from the marginal distribution of $\hat{\theta}_i$'s, and inferences are then based on the estimated conditional (or posterior) distribution of θ_i, $f(\theta_i \mid \hat{\theta}_i, \hat{\boldsymbol{\beta}}, \hat{\sigma}_v^2)$. In particular, the mean of the estimated posterior distribution is the EB estimator $\tilde{\theta}_i^{\text{EB}}$. Under normality, $\tilde{\theta}_i^{\text{EB}}$ is identical to the EBLUP estimator $\tilde{\theta}_i$, but the EB approach is applicable generally for any joint distribution. It should be noted that the EB approach is essentially frequentist because it uses only the sampling model and the linking model which can be validated from the data; no priors on the model parameters are involved unlike in the HB approach.

As a measure of variability of $\tilde{\theta}_i^{\text{EB}}$, the variance of the estimated posterior is used. Under normality, it is given by $g_{1i}(\hat{\sigma}_v^2) = \hat{\gamma}_i\psi_i$ which leads to severe underestimation of true variability as measured by MSE. Laird and Louis (1987) proposed a parametric bootstrap method to account for the variability in $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_v^2$, but Butar and Lahiri (1997) showed that it is not second-order correct, *i.e.*, its bias involves terms of order m^{-1}, unlike the bias of (3.7) or (3.8). By correcting this bias, they obtained an estimator which is identical to the area-specific MSE estimator (3.8). Therefore, corrected EB and EBLUP lead to the same result under normality.

3.3 HB method

The HB approach has been extensively used for small area estimation. It is straightforward, inferences are "exact" and it can handle complex problems using recently developed Monte Carlo Markov Chain (MCMC) methods, such as the Gibbs sampler. A prior distribution on the model parameters (also called hyper parameters) is specified and the posterior distribution of the small area totals Y_i or $g(Y_i) = \theta_i$ is then obtained. Inferences are based on the posterior distribution; in particular, Y_i or θ_i is estimated by its posterior mean and its precision is measured by its posterior variance.

For the basic area level model, (2.1) and (2.2), with normality of v_i and e_i, the posterior mean $E(\theta_i \mid \hat{\boldsymbol{\theta}})$ and the posterior variance $V(\theta_i \mid \hat{\boldsymbol{\theta}})$ are obtained in two stages, where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_m)'$. In the first stage, we obtain $E(\theta_i \mid \hat{\boldsymbol{\theta}}, \sigma_v^2)$ and $V(\theta_i \mid \hat{\boldsymbol{\theta}}, \sigma_v^2)$ for fixed σ_v^2, assuming an improper prior, $f(\boldsymbol{\beta}) \propto \text{const.}$, on $\boldsymbol{\beta}$ to reflect absence of prior information on $\boldsymbol{\beta}$. The conditional posterior mean, given σ_v^2, is identical to the BLUP estimator $\tilde{\boldsymbol{\theta}}_i(\sigma_v^2)$ and the conditional posterior variance is equal to $g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)$. At the second stage, we take account of the uncertainty about σ_v^2 by first calculating its posterior distribution $f(\sigma_v^2 \mid \hat{\boldsymbol{\theta}})$, assuming a prior distribution on σ_v^2 and prior independence of $\boldsymbol{\beta}$ and σ_v^2. The posterior mean and variance are then obtained as

$$\tilde{\theta}_i^{\text{HB}} = E(\theta_i \mid \hat{\boldsymbol{\theta}}) = E_{\sigma_v^2 \mid \hat{\boldsymbol{\theta}}}[\tilde{\theta}_i(\sigma_v^2)] \qquad (3.9)$$

$$V(\theta_i \mid \hat{\boldsymbol{\theta}}) = E_{\sigma_v^2 \mid \hat{\boldsymbol{\theta}}}[g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)] + V_{\sigma_v^2 \mid \hat{\boldsymbol{\theta}}}[\tilde{\theta}_i(\sigma_v^2)] \qquad (3.10)$$

where $E_{\sigma_v^2|\hat{\boldsymbol{\theta}}}$ and $V_{\sigma_v^2|\hat{\boldsymbol{\theta}}}$ denote the expectation and variance with respect to $f(\sigma_v^2|\hat{\boldsymbol{\theta}})$. No closed form expressions for (3.9) and (3.10) exist, but in this simple case they can be evaluated numerically using only one-dimensional integration. For complex models, high-dimensional integration is often involved and it is necessary to use MCMC-type methods to overcome the computational difficulties.

It follows from (3.9) that $\tilde{\theta}_i^{\,\mathrm{HB}} \approx \tilde{\theta}_i(\hat{\sigma}_v^2)$ but (3.10) shows that ignoring uncertainty about σ_v^2 and using $g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2)$ as a measure of variability can lead to significant underestimation.

If the assumed prior $f(\sigma_v^2)$ is proper and informative, the HB approach encounters no difficulties. On the other hand, an improper prior $f(\sigma_v^2)$ could lead to an improper posterior (Hobert and Casella 1996). In the latter case, we cannot avoid the difficulty by choosing a diffuse proper prior on σ_v^2 because we will be simply approximating an improper posterior by a proper posterior.

To illustrate the use of Gibbs sampling, we again consider the basic area level model under normality. To implement Gibbs sampling assuming the prior $f(\tau_v = \sigma_v^{-2})$ is a gamma(a, b), $a > 0, b > 0$, we need the following Gibbs-conditional distributions:

$$\text{(i)} \quad \boldsymbol{\beta}\,|\,\boldsymbol{\theta}, \sigma_v^2, \hat{\boldsymbol{\theta}} \sim N_p[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\theta}, \sigma_v^2(\mathbf{X}'\mathbf{X})^{-1}] \quad (3.11)$$

$$\text{(ii)} \quad \theta_i\,|\,\boldsymbol{\beta}, \sigma_v^2, \hat{\boldsymbol{\theta}} \overset{\text{ind}}{\sim} N(\tilde{\theta}_i(\boldsymbol{\beta}, \sigma_v^2)) = \gamma_i\theta_i + (1-\gamma)\mathbf{x}_i'\boldsymbol{\beta}, \gamma_i\psi_i), \; i = 1, ..., m \quad (3.12)$$

$$\text{(iii)} \quad \sigma_v^{-2}\,|\,\boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \sim G\left[\frac{m}{2} + a, \frac{1}{2}\sum(\theta_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + b\right], \quad (3.13)$$

where $\mathbf{X}$ is the $m \times p$ matrix with $\mathbf{x}_i'$ as the i^{th} row and $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)'$. The Gibbs algorithm is as follows: (a) Using starting values $\theta_i^{(0)}$ and $\sigma_v^{2(0)}$ draw $\boldsymbol{\beta}^{(1)}$ from (3.11). (b) Draw $\theta_i^{(1)}$, $i = 1, ..., m$ from (3.12) using $\boldsymbol{\beta}^{(1)}$ and $\sigma_v^{2(0)}$. (c) Draw $\sigma_v^{2(1)}$ from (3.13) using $\theta_i^{(1)}$, $i = 1, ..., m$ and $\boldsymbol{\beta}^{(1)}$. Steps (a)-(c) complete one cycle. Perform a large number of cycles, say t, called "burn-in period", until convergence and then treat $(\boldsymbol{\beta}^{(t+j)}, \sigma_v^{2(t+j)}, \theta_i^{(t+j)}, j = 1, ..., J)$ as J samples from the joint posterior of $\boldsymbol{\beta}, \sigma_v^2$ and $\theta_i, i = 1, ..., m$. Other methods use multiple parallel runs instead of a single long run as above. Parallel runs can be wasteful because initial "burn-in" periods are discarded from each run. But a single long run may leave a significant portion of the space generated by the joint posterior unexplored.

The posterior mean and the posterior variance are estimated as

$$\tilde{\theta}_i^{\mathrm{HB}} \approx \frac{1}{J}\sum_j \tilde{\theta}_i[\sigma_v^{2(t+j)}] = \frac{1}{J}\sum_j \tilde{\theta}_i(j) = \tilde{\theta}_i(\cdot) \quad (3.14)$$

and

$$V(\theta_i\,|\,\hat{\boldsymbol{\theta}}) \approx \frac{1}{J}\sum_j [g_{1i}(\sigma_v^{2(t+j)}) + g_{2i}(\sigma_v^{2(t+j)})] + \frac{1}{J}\sum_j [\tilde{\theta}_i(j) = \tilde{\theta}_i(\cdot)]^2. \quad (3.15)$$

The estimator $\tilde{\theta}_i(\cdot)$ has smaller simulation error than the estimator is a conditional expectation and the well-known $J^{-1}\sum_j \theta_i^{(t+j)}$ because $\tilde{\theta}_i(\sigma_v^2)$ is a conditional expectation and the well-known Rao-Blackwell theorem holds. It is therefore advisable to do analytical calculations first before applying Gibbs sampling.

For the basic area level model, all the conditional distributions, (3.11)-(3.13), are in a closed form and, therefore, samples can be generated directly. But for more complex models, some of the conditionals may not have closed form in which case alternative algorithms, such as Metropolis-Hastings within Gibbs, are needed to draw samples from the joint posterior distribution. We refer the reader to Brooks (1998) for an excellent review of the MCMC methods. Software, called BUGS and CODA, are readily available for implementing MCMC and convergence diagnostics, but caution should be exercised in using MCMC methods. For example, Hobert and Casella (1996) demonstrated that the Gibbs sampler could lead to seemingly reasonable inferences about a nonexistent posterior distribution. This happens when the posterior is improper and yet all the Gibbs-conditional distributions are proper. Another difficulty with MCMC is that the convergence diagnostics tools can fail to detect the sorts of convergence failure that they were designed to identify (Cowles and Carlin 1996). Further difficulties include the choices of t for the burn-in period, number of simulated samples, J, and the starting values.

3.4 Some recent applications

(1) Dick (1995) used the basic area level model (2.3) to estimate net under coverage rates in the 1991 Canadian Census. The goal is to estimate 96 adjustment factors $\theta_i = T_i/C_i$, corresponding to 2(sex) × 4(age) × 12(province) combinations, where T_i is the true (unknown) count and C_i is the census count in the i^{th} area domain the net undercoverage rate in the i^{th} area is given by $U_i = 1 - \theta_i^{-1}$. Direct estimates $\hat{\theta}_i$ were obtained from a post enumeration survey, and sampling variances ψ_i were derived through smoothing of estimated variances, assuming ψ_i is proportional to some power of C_i. Explanatory variables, $\mathbf{x}$, were selected from a set of 42 variables by backward stepwise regression. EBLUP (EB) estimates of θ_i were used and their MSE estimated using (3.7) with REML estimate of σ_v^2. The EB adjustment factors $\tilde{\theta}_i^{\mathrm{HB}}$ were converted to estimates of missing persons, $M_i = T_i - C_i$, and these estimates were raked to ensure consistency with direct estimates of marginal totals. The raked EB estimates, $\tilde{\theta}_i^{\mathrm{R}}$ were used as the final

estimates of M_i's. MSE estimate of $\tilde{\theta}_i^R$ was obtained as $[\text{mse}(\tilde{\theta}_i^{HB})](\tilde{\theta}_i^R / \tilde{\theta}_i^{HB})^2$. This somewhat *ad hoc* method ensures that the coefficient of variation (CV) of $\tilde{\theta}_i^{HB}$ is retained by $\tilde{\theta}_i^R$, but properties of this method remains to be investigated.

(2) The basic area level model (2.3) with $\theta_i = \log Y_i$ has been recently used to produce model-based county estimates of poor school-age children in U.S.A. (Fisher and Siegel 1997; National Research Council 1998). Using these estimates, the US Department of Education allocates over 7 billion of federal funds annually to counties. The difficulty with unknown ψ_i was handled by using a model of the form (2.3) for the census year 1990, for which reliable estimates $\hat{\psi}_{ic}$ of sampling variances, ψ_{ic}, are available and assuming the census small area effects v_{ic} follow the same distribution as v_i, *i.e.*, $N(0, \sigma_v^2)$. Under the latter assumption, an estimate of σ_v^2 was obtained from the census data assuming $\hat{\psi}_{ic} = \psi_{ic}$ and used in the current model (2.3), assuming $\psi_i = \sigma_e^2 / n_i$, to get an estimate of σ_e^2. The resulting estimate, $\tilde{\psi}_i = \tilde{\sigma}_e^2 / n_i$, was treated as the true ψ_i in developing EBLUP estimates, $\tilde{\theta}_i$, of θ_i. The small area (county) totals Y_i (number of school-age children in poverty) can then be estimated as $\tilde{Y}_i = \exp(\tilde{\theta}_i)$, but a more refined method based on the mean of lognormal distribution was used: $\tilde{\tilde{Y}}_i = \exp\{\tilde{\theta}_i + 1/2\,\text{MSE}(\tilde{\theta}_i)\}$, ignoring the g_{3i}-term in (3.7) which was found to be small. The MSE of $\tilde{\tilde{Y}}_i$ was estimated using the approximation $\text{MSE}(\tilde{\theta}_i) \approx \text{CV}^2(\tilde{\tilde{Y}}_i)$. The estimates $\tilde{\tilde{Y}}_i$ were raked to agree with model-based state estimates obtained from a state model. The reader is referred to National Research Council (1998) for details on x-variables used in the county model and evaluation of the models. Several criteria were used for evaluating the models and the estimates, including regression diagnostics and comparisons to the 1990 Census counts.

(3) Other applications of the basic area level model include the following: (i) Estimation of unemployment rates at census tract level (Chand and Alexander 1995); (ii) Estimation of counts in employment categories and household income categories at the Congressional District level (Griffiths 1996); (iii) Estimation at the provincial level in the Italian Labour Force Survey (Falorsi, Falorsi and Russo 1995).

4. Extensions

We now present some recent extensions and applications of the basic area level model in section 4.1 and those of the basic unit level model in section 4.2.

4.1 Area-level models

Recent extensions of the basic area level model include multivariate and time series models and models for disease mapping, as noted in section 2.

4.1.1 Multivariate models

Datta, Ghosh, Nangia and Natarajan (1996) used multivariate area level (Fay-Herriot) models to develop HB estimators of median income of four-person families for U.S. states. Here $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \theta_{i3})'$ with θ_{i1}, θ_{i2} and θ_{i3} denoting the true median incomes of four-, three- and five-person families in state i. Adjusted census median income and base- year census median income for the three groups were used as explanatory variables. Diffuse priors on model parameters were used along with Gibbs sampling. The resulting HB estimators, HB^3, for four-person families in 1979 were compared to the direct Current Population Survey (CPS) estimators and univariate and bivariate model-based HB estimators, HB^1 and HB^2, treating the 1979 estimates, available from the 1980 census data, as the true values. In terms of relative absolute error averaged over the states, the three HB estimators performed similarly, but outperformed the direct CPS estimates. In this application, the univariate estimator HB^1 worked well and it is not necessary to use more complicated estimators based on multivariate models. Estimates of θ_{i1} are used for administering an energy assistance program to low-income families.

Longford (1999) obtained multivariate shrinkage (composite) estimators of small area means and proportions, and illustrated their superiority over univariate shrinkage estimators.

4.1.2 Time series models

(1) Ghosh *et al.* (1996) developed HB estimators under the time series linking model given by (2.5) and (2.6) and applied them to estimate median income of four person families using direct estimates $\hat{\theta}_{it}$, $i = 1, ..., 51$; $t = 1, ..., 10$ for the 51 states over a ten year period.

(2) Datta *et al.* (1994) used the time series model (2.10) with $\mathbf{u}_t$ following (2.6) and developed HB estimators. They also used methods for validating the model, based on cross-validation. They applied the methods to estimate monthly unemployment rates for U.S. states. HB estimators performed significantly better than the CPS estimates, as measured by the CPS and HB standard errors. We refer the reader to Datta *et al.* (1994) for details on the x-variables used. Datta, Lahiri, Maiti, and Lu (1999) used the linking model (2.8) with a random walk model on the u_{it}'s, but added extra terms to (2.8) to reflect seasonal variation in unemployment rates.

(3) Datta, Lahiri and Maiti (1999) and You (1999) obtained EBLUP (EB) estimators and associated second-order correct estimators of MSE for the time

series/cross-sectional linking model (2.8) with a random walk model on u_{it}'s. Datta *et al.* used ML and REML estimators of model parameters while You employed the method of moments estimators.

Datta, Lahiri and Maiti (1999) used EB estimators to estimate median income of four-person families by U.S. states using time series and cross-sectional data. They employed the linking model (2.8) with a random walk model on u_{it}'s. Using the 1979 estimates available from the 1980 Census data as the true values, they compared the EB (EBLUP) estimates with the HB estimates of Ghosh *et al.* (1996) and the CPS direct estimates. In terms of absolute relative bias averaged over states, EB performed better than HB and both EB and HB performed much better than the CPS direct estimate. In terms of coefficient of variation, EB again performed better than HB and CPS; second-order correct estimate of MSE of EB was used.

4.1.3 Disease mapping models

Maiti (1998) used the model $y_i|\theta_i \overset{\text{ind}}{\sim} P(n_i\theta_i)$ and $\beta_i = \log\theta_i \overset{\text{i.i.d.}}{\sim} N(\mu,\sigma^2)$ with diffuse prior on μ and a gamma prior on σ^{-2}. He obtained HB estimators of θ_i, and the posterior variance of θ_i, and applied the results to the well-known lip cancer data from Scottish Counties (small areas); see Clayton and Kaldor (1987) for details. He also studied HB estimation under the spatial dependence model for β_i's mentioned in section 2.1. Estimates of θ_i's are very similar for both the models but standard errors for the spatial model are smaller than those under the first model. Lahiri and Maiti (1996) obtained EB estimators and second order correct estimators of MSE under the Clayton-Kaldor model mentioned in section 2.1, and illustrated the method on the Clayton-Kaldor data set. Nandram *et al.* (1998) used the age-group specific models, mentioned in section 2.1, to obtain HB estimators and also developed Bayesian methods to compare alternative models, using three different measures of fit. They applied the results to estimate age specific and age adjusted mortality rates for Health Services Area's (sets of counties based on where residents seek routine hospital care) for the disease category "all cancers for white males".

4.1.4 Other extensions

Datta and Lahiri (1995) considered robust HB estimation using a class of scale mixtures of normal distributions on the random effects v_i with the basic area level model. This class includes t, Laplace and logistic distributions; Cauchy distribution for outlier areas was adopted.

You (1999) considered the more realistic sampling model (2.4) on $\hat{Y}_i$ with sampling errors e_i^* and the linking model (2.1). Assuming $V(e_i^*|Y_i) = \psi_i^2 Y_i^2$ and $\hat{\theta}_i = \log(\hat{Y}_i)$, he used HB methods to demonstrate that for small sample sizes the posterior inferences under the sampling model (2.4) can be significantly different from those under the sampling model on $\hat{\theta}_i$.

4.2 Unit level models

Recent extensions at the basic unit level model include multivariate models, two-way and two-level models, random error variance models and logistic linear mixed models, as noted in section 2.

4.2.1 Nested error regression models

Rao and Choudhry (1995) provided an overview of small area estimation in the context of business surveys. They also studied the performance of EBLUP estimator of a small area total relative to traditional estimators through simulation using real and synthetic populations.

As noted in section 2, model-based estimators for unit level models do not depend on the survey weights. Prasad and Rao (1999) obtained model-assisted estimators for the nested error regression model that depend on survey weights $\tilde{w}_{ij}$ and remain design-consistent as the sample size, n_i, increases. The unit level sample model is first reduced to

$$\bar{y}_{iw} = \bar{\mathbf{x}}_{iw}'\boldsymbol{\beta} + v_i + \bar{e}_{iw}, \tag{4.1}$$

where $\bar{y}_{iw} = \sum_j w_{ij} y_{ij}$ with $w_{ij} = \tilde{w}_{ij}/\sum_j \tilde{w}_{ij}$ and similar expressions for $\bar{x}_{iw}$ and $\bar{e}_{iw}$. A pseudo-BLUP estimator of $\theta_i = \bar{\mathbf{X}}_i'\boldsymbol{\beta} + v_i$, for fixed σ_v^2 and σ_e^2, say $\hat{\theta}_{iw}(\sigma_v^2, \sigma_e^2)$ is then obtained from the reduced model (4.1), noting that $\bar{e}_{iw} \overset{\text{ind}}{\sim} N(0, \sigma_e^2 \sum_j w_{ij}^2)$, where $\bar{\mathbf{X}}_i$ is the vector of known population means and $\bar{Y}_i \approx \theta_i$ for large N_i (This estimator is called pseudo-BLUP because it is different from the BLUP estimator under the full unit-level sampling model). The unknown parameters σ_v^2 and σ_e^2 are then replaced by model-consistent estimators $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ under the full model to obtain the pseudo-EBLUP estimator $\hat{\theta}_{iw} = \hat{\theta}_{iw}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$. This estimator is model-assisted and it is approximately design and model unbiased even if the sample design is nonignorable. Prasad and Rao (1999) also obtained a second-order correct estimator of $\text{MSE}(\hat{\theta}_{iw})$. You and Rao (1999b) developed a pseudo-HB methodology which leads to estimators similar to the pseudo-EBLUP estimators of Prasad and Rao (1999).

Singh, Stukel and Pfeffermann (1998) made a comparison of frequentist and Bayesian measures of error, using analytical and empirical methods for the basic unit-level model.

Stukel and Rao (1999) obtained EBLUP estimators and associated approximately unbiased (or second-order) correct MSE estimators under two-way nested error regression models. Simulation results of Stukel and Rao (1999) suggested that the behaviour of relative bias of MSE estimators is more complex than in the one-way case.

4.2.2 Two-level models

Moura and Holt (1999) obtained EBLUP estimators and associated second-order correct MSE estimators for the two-

level models. They obtained EBLUP estimators and used them on data from a sample of 951 retail stores in Southern Brazil classified into 73 small areas. They compared the average second order correct MSE of the estimators to the average MSE value for the nested error regression model to demonstrate improvement in efficiency. You and Rao (1999a) applied HB methods to the Brazilian data. They studied three different two level models: (1) equal error variances; (2) unequal error variances; (3) random error variances. Bayesian diagnostics revealed that model (2) fits the data better than models (1) and (3).

4.2.3 Random on error variances models

Arora and Lahiri (1997) studied the unit-level model with random error variances σ_i^2 and assumed $\sigma_i^{-2} \overset{\text{i.i.d.}}{\sim} G(a, b)$. They obtained the EB estimator of small area mean $\overline{Y}_i$ and applied the Laird-Louis bootstrap to estimate its MSE, taking account of the variability due to estimation of model parameters.

Arora and Lahiri (1997) obtained a reduced model from the unit level random error variances model by incorporating survey weights. They performed HB analysis on the reduced model with $\sigma_i^{-2} \overset{\text{i.i.d.}}{\sim} G(a, b)$, and applied the results to estimate the average weekly consumer expenditures of various items, goods and services for $m = 43$ publication areas (small areas) in U.S.A.

4.2.4 General linear mixed models

Datta and Lahiri (1997) studied the general linear mixed model with a block diagonal covariance structure, (2.12). They developed EBLUP estimators and associated second-order correct estimators of MSE, using REML or ML estimators. In the case of ML estimators an extra term of order $O(m^{-1})$ should be subtracted. Das, Rao et You (1999) extended these results to the general mixed ANOVA model (2.13) in which case the asymptotic set-up is more complex.

4.2.5 Multivariate nested error regression models

Datta, Day and Basawa (1999) obtained EBLUP (EB) estimators and second order correct estimators of MSE, for the multivariate nested error regression models. They conducted a simulation study using the sample sizes and auxiliary variable values given by Battese, Harter and Fuller (1988). Further, they estimated the model parameters for their multivariate model using Battese *et al.*, data on crop areas under corn and soybeans for $m = 12$ counties in North-Central Iowa. Treating the estimated parameters as true values, they generated simulated samples and showed that the multivariate approach can achieve substantial improvement over the univariate approach inefficiency.

4.2.6 Logistic linear mixed models

Farrell, MacGibbon and Tomberlin (1997a, 1997b) studied EB estimation for binary y, assuming the sampling model $y_{ij} | \theta_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij})$ and the linking logistic model $\log\{\theta_{ij}/(1-\theta_{ij})\} = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i$ with $v_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2)$. The conditional distribution of θ_{ij}'s is approximated by a multivariate normal to get an EB estimator of local area proportion $\overline{Y}_i$. They employed the bootstrap method of Laird and Louis (1987) to get a bootstrap-adjusted estimate of variability associated with the EB estimator. But results of Butar and Lahiri (1997) for the linear case suggest that the bootstrap method may not be second-order correct in the nonlinear case as well. Jiang and Lahiri (1998) also studied EB estimation for the above model and obtained the EB estimator exactly through one-dimensional numerical integration. They called the EB estimator an empirical best predictor (EBP) which may be more appropriate because no priors on model parameters are involved. Employing method of moment estimators of model parameters $\boldsymbol{\beta}$ and σ_v^2, they also obtained an approximation to MSE of the EB estimator correct to terms of order m^{-1}. Jiang, Lahiri, and Wan (1999) proposed a jackknife method of estimating MSE that is applicable to general longitudinal linear and generalized linear mixed models. This method leads to second-order correct MSE estimators and looks promising. But one needs to recompute the REML estimates of model parameters by deleting each area in turn. The computations can be significantly reduced by using a single step of the Newton-Raphson algorithm with the estimates from the full sample as starting values. Properties of this simplification remain to be studied. Booth and Hobert (1998) argued that the conditional MSE of the EBP given the i^{th} area data is more relevant as a measure of variability than the unconditional MSE because it is area-specific. Fuller (1989) earlier proposed a similar criterion in the context of linear mixed models. But the MSE estimator (3.8) shows that it is possible to obtain area-specific estimators of the unconditional MSE, at least in the linear model case. Also, it is not clear how one should proceed with the conditioning when two or more small area estimators need to be aggregated to obtain an estimator for a larger area. How would one define the conditional MSE of the larger area estimator?

Malec *et al.* (1997) used logistic linear mixed models and the HB approach to estimate proportions for demographic groups within U.S. states. Data from the National Health Interview Survey were used for this purpose. Cross-validation methods were used to evaluate the model fit. For one of the binary variables observed for respondents to the 1990 census long form, they compared the estimates from alternative methods and models with the very accurate census estimates of true values. For logistic linear mixed models, not all the conditional distributions for Gibbs sampling have closed form unlike those obtained for the probit linear mixed model derived from a latent variable approach (Das *et al.* 1999).

Malec, Davis and Cao (1996, 1999) studied logistic linear mixed models to estimate overweight prevalence for subgroups (small areas) using National Health and Nutrition Examination Survey (NHANES III) data. Again, HB

methods were used but survey weights were incorporated using a pseudo-likelihood. Folsom, Shah and Vaish (1999) studied general logistic mixed linear models in the context of estimating substance abuse in U.S. states from the 1994-1996 National Household Surveys on Drug Abuse. They developed survey-weighted pseudo HB estimators and associated posterior variance, using MCMC methods.

Ghosh *et al.* (1998) applied the HB approach to generalized linear mixed models and used the results on two real data sets. The first data set, based on a 1991 sample of all persons in 15 geographical regions of Canada consists of responses classified into four categories to the question "Have you experienced any negative impact of exposure to health hazards in the work place?" Objective here is to estimate the proportion of workers in each of the four response categories for every one of 60 groups cross-classified by 16 geographical regions and 4 demographic (age × sex) groups. The second data set relates to cancer mortality rates for the 115 counties in Missouri during 1972-81.

5. Discussion

We briefly discussed, in section 1, survey design issues that have an impact on small area statistics. Preventive measures at the design stage, such as those proposed by Singh, Gambino and Mantel (1994), may reduce the need for indirect estimators significantly, although for many applications sample sizes in some domains of interest may not be large enough to provide adequate precision even after taking such measures. As noted in section 1, sometimes the survey is deliberately designed to oversample specific domains at the expense of small samples or even no samples in other domains (areas) of interest.

We have provided a brief overview of the literature, over the past five years or so, on model-based small area estimation. The methodological developments and applications are both impressive, but it is necessary to exercise caution in using model-based methods because of the underlying assumptions. Good auxiliary information related to the variables of interest plays a vital role in model-based inference. As noted by Schaible (1996), expanded access to auxiliary information through coordination and cooperation among federal agencies is needed.

Model validation also plays an important role in model-based estimation. Fay and Herriot (1979), Ghosh and Rao (1994), Dick (1995), Malec *et al.* (1997), Datta, Lahiri, Maiti, and Lu (1999), You and Rao (1999a), National Research Council (1998) and others used some methods for model validation and illustrated their application. But the available methods for handling models with random effects are not as extensive as those used for the standard linear and non-linear models with only fixed effects. More work, both classical and Bayesian, on model diagnostics for random effects models is needed.

Area-level models have wider scope than the unit level models because area-level auxiliary information is more readily available than unit-level auxiliary data. But the assumption of known sampling variances, ψ_i, is quite restrictive, although the methods used in the applications (section 3.4) seem to be promising. It should be noted that errors in estimating ψ_i do not affect the model-unbiasedness of the EBLUP(EB) estimators provided the mean of θ_i in the linking model (2.1) is correctly specified. But the efficiency of the estimator is affected as well as the validity of the MSE estimators (3.7) and (3.8). More work is needed on obtaining good approximations to the sampling variances. This task becomes more difficult when using multivariate and time series area levels models because sampling covariances are also needed.

Recent work on incorporating survey weights into model-based estimation under unit-level models is promising, but the assumption that the sample design is ignorable may not be true for some applications. Krieger and Pfeffermann (1997) proposed methods for direct estimation of large area parameters that take account of the sample selection effects. It would be useful to extend this work to indirect estimation of small area parameters.

The Hierarchical Bayes (HB) approach is a powerful method for small area estimation because it can handle complex problems and the inferences are "exact". But, as noted in section 3.3, caution should be exercised in the choice of improper prior distributions on the model parameters.

We studied model-based estimates of small area totals or means, but they may not be suitable if the objective is to identify domains with extreme population values or to rank domains or to identify domains that fall below or above some prespecified level. Ghosh and Rao (1994) reviewed some methods for handling the latter cases. For a simple model, Shen and Louis (1998) proposed "triple-goal" estimators that produce good ranks, a good distribution and good area – specific estimators. It would be useful to extend their approach to handle more complex models that are suitable for small area estimation.

Acknowledgements

This work was supported by a research grant from the Natural Sciences and Engineering Research Council of Canada. I am thankful to Dr. Graham Kalton and the Editor for constructive suggestions.

References

Arora, V., and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Booth, J.G., and Hobert, J.P. (1998). Standard errors of predictors in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 362-372.

Brooks, S.P. (1998). Markov Chain Monte Carlo method and its application. *The Statistician*, 47, 69-100.

Butar, P.B., and Lahiri, P. (1997). On the Measures of Uncertainty of Empirical Bayes Small Area Estimators. Technical Report, Department of Mathematics and Statistics, University of Nebraska-Lincoln.

Chand, N., and Alexander, C.H. (1995). Indirect estimation of rates and proportions for small areas with continuous measurement. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 549-554.

Chaudhuri, A., and Adhikary, A.K. (1995). On generalized regression estimators of small domain totals – an evaluation study. *Pakistan Journal of Statistics*, 11, 173-189.

Clayton, D., and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671-681.

Cowles, M.K., and Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883-904.

Das, K., and Rao, J.N.K. (1999). Second order approximations for standard errors of empirical BLUP estimators in general mixed ANOVA models. Paper under preparation.

Das, K., Rao, J.N.K. and You, Y. (1999). Small Area Estimation for Binary Variables Using Probit Linear Mixed Models. Paper under preparation.

Datta, G.S., Day, B. and Basawa, I. (1999). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference*, 75, 269-279.

Datta, G.S., Ghosh, M., Nangia, N. and Natarajan, K. (1996). Estimation of median income of four-person families: A Bayesian approach. In: *Bayesian Analysis in Statistics and Econometrics*, (Eds., D.A. Berry, K.M. Chaloner, and J.K. Geweke). New York: John Wiley & Sons, Inc., 129-140.

Datta, G.S., and Lahiri, P. (1995). Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis*, 54, 310-328.

Datta, G.S., and Lahiri, P. (1997). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictor in Small-Area Estimation Problems. Technical Report, Department of Statistics, University of Georgia-Athens.

Datta, G.S., Lahiri, P. and Lu, K.L. (1994). Hierarchical Bayes Time Series Modeling in Small Area Estimation With Applications. Technical Report, Department of Mathematics and Statistics, University of Nebraska-Lincoln.

Datta, G.S., Lahiri, P. and Maiti, T. (1999). Empirical Bayes Estimation of Median Income of Four Person Families by States Using Time Series and Cross-Sectional Data. Technical Report, Department of Statistics, University of Georgia-Athens.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the U.S. states. *Journal of the American Statistical Association*, 94, 1074-1082.

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

Falorsi, P.P., Falorsi, S. and Russo, A. (1994). Empirical comparison of small area estimation methods for the Italian labour force survey. *Survey Methodology*, 20, 171-176.

Falorsi, P.D., Falorsi, S. and Russo, A. (1995). Small area estimation at provincial level in the Italian Labour Force Survey. *Proceedings of the 1995 Annual Research Conference*, *U.S. Bureau of the Census*, 617-635.

Farrell, P.J., MacGibbon, B. and Tomberlin, T.J. (1997a). Empirical Bayes estimates of small area proportions in multistage designs. *Statistica Sinica*, 7, 1065-1083.

Farrell, P.J., MacGibbon, B. and Tomberlin, T.J. (1997b). Empirical Bayes small area estimation using logistic regression models and summary statistics. *Journal of Business and Economic Statistics*, 15, 101-108.

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.

Fisher, R., and Siegel, P. (1997). Methods used for small area poverty and income estimation. *Proceedings of the Social Statistics Section*, American Statistical Association.

Folsom, R., Shah, B. and Vaish, A. (1999). Substance Abuse in States: Model Based Estimates from the 1994-1996 National Household Surveys on Drug Abuse. Methodology report. Research Triangle Institute.

Fuller, W.A. (1989). Prediction of True Values for the Measurement Error Model. Paper presented at the Conference on Statistical Analysis of Measurement Error Models, Humboldt State University.

Fuller, W.A., and Harter, R.M. (1987). The multivariate components of variance model for small area estimation. In: *Small Area Statistics* (Eds., R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh). New York: John Wiley & Sons, Inc., 103-123.

Ghosh, M. Nangia, N. and Kim, D.H. (1996). Estimation of median income of four-person families: A Bayesian approach. *Journal of the American Statistical Association*, 91, 1423-1431.

Ghosh, M., Natarajan, K., Kim, D. and Walker, L.A. (1997). Hierarchical Bayes GLM's for the Analysis of Spatial Data: An Application to Disease Mapping. Technical report, University of Florida-Gainsville.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Griffiths, R. (1996). Current Population Survey Small Area Estimation for Congressional Districts. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 314-319.

Hobert, J.P., and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461-1479.

Jiang, J. (1996). REML estimation: Asymptotic behaviour and related topics. *Annals of Statistics*, 24, 255-286.

Jiang, J., and Lahiri, P. (1998). Empirical Best Prediction for Small Area Inference With Binary Data. Technical report, Department of Mathematics and Statistics, University of Nebraska- Lincoln.

Jiang, P., Lahiri, P. and Wan, S. (1999). Jackknifing the Mean Squared Error of Empirical Best Predictor. Technical report, Department of Statistics, Case Western Reserve University.

Jiang, J., Lahiri, P. and Wu, C. (1998). On Pearson- χ^2 Testing With Unobservable Frequencies and Mixed Model Diagnostics. Technical report, Department of Statistics, Case Western Reserve University.

Kleffe, J., and Rao, J.N.K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under random error variance linear model. *Journal of Multivariate Analysis*, 43, 1-15.

Krieger, A.M., and Pfeffermann, D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics,* 13, 123-142.

Lahiri, P.A., and Maiti, T. (1996). Empirical Bayes Estimation of Mortality From Diseases for Small Area. Technical report, Department of Mathematics and Statistics, University of Nebraska-Lincoln.

Lahiri, P.A., and Rao, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 82, 758-766.

Laird, N.M., and Louis, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82, 739-750.

Longford, N.T. (1999). Multivariate shrinkage estimations small area means and proportions. *Journal of the Royal Statistical Society*, Series A, 182, 227-245.

Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.

Malec, D., Davis, W.W. and Cao, X. (1996). Small area estimates overweight prevalence using the third National Health and Nutrition Examination Survey (NHANES III). *Proceedings of the Section on Survey Research Method*, American Statistical Association, 326-331.

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.

Malec, D., Sedransk, J., Moriarity, C.L. and Leclere, F. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association*, 92, 815-826.

Marker, D.A. (1999). Organization of small area estimators using generalized linear regression framework. *Journal of Official Statistics,* 15, 1-24.

Moura, F., and Holt, D. (1999). Small area estimation using multi level models. *Survey Methodology*, 25, 73-80.

Nandram, B., Sedransk, J. and Rickle, L. (1998). Regression Analysis of Mortality Rates for U.S. Health Service Areas. Technical report, Worcester Polytechnic Institute.

National Research Council (1998). *Small Area Estimation of School-Age Children in Poverty.* Interim Rep. 2. Washington, D.C.: National Research Council.

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.

Prasad, N.G.N., and Rao, J.N.K. (1999). On robust estimation using a simple random effects model. *Survey Methodology*, 25, 67-72.

Raghunathan, T.E. (1993). A quasi-empirical Bayes method for small area estimation. *Journal of the American Statistical Association*, 88, 1444-1448.

Rao, J.N.K. (1998). EB and EBLUP in Small Area Estimation. Technical report, Laboratory for Research in Statistics and Probability, Carleton University.

Rao, J.N.K., and Choudhry, G.H. (1995). Small area estimation: Overview and empirical study. In: *Business Survey Methods* (Eds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott). New York: John Wiley & Sons, Inc., 527-542.

Rao, J.N.K., and Yu, M. (1992). Small area estimation by combining time series and cross-sectional data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1-9.

Rao, J.N.K., and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22, 511-28.

Rivest, L.P., and Belmonte, E. (1999). The Conditional Mean Squared Errors of Small Area Estimators in Survey Sampling. Technical report, Laval University.

Schaible, W.L. (Editor) (1996). *Indirect Estimators in U.S. Federal Programs. Lecture Notes in Statistics No. 108*. New York: Springer-Verlog.

Shen, W., and Louis, T.A. (1998). Triple-goal estimators in two-stage hierarchical models. *Journal of the Royal Statistical Society*, Series B, 60, 455-471.

Singh, A.C., Mantel, H.J. and Thomas, B.W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology*, 20, 33-43.

Singh, A.C., Stukel, D.M. and Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, Series B, 60, 377-396.

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-22.

Stukel, D.M., and Rao, J.N.K. (1999). On small-area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference*, 78, 131-147.

You, Y. (1999). Hierarchical Bayes and Related Methods for Model-Based Small Area Estimation. Ph.D. Thesis, School of Mathematics and Statistics, Carleton University.

You, Y., and Rao, J.N.K. (1999a). Hierarchical Bayes estimation of small area means using multi-level models. Invited paper, *Proceedings of the IASS Satellite Conference on Small Area Estimation,* Riga, Latvia, 171-185.

You, Y., and Rao, J.N.K. (1999b). Pseudo hierarchical Bayes small area estimation using sampling weights. *1999 Proceedings of the Survey Methods Section*, *Statistical Society of Canada*, in press.