Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics
doi: 10.1111/sjos.12061

Prediction Error of Small Area Predictors Shrinking Both Means and Variances

TAPABRATA MAITI
Department of Statistics and Probability, Michigan State University

HAO REN
CTB/McGraw-Hill

SAMIRAN SINHA
Department of Statistics, Texas A&M University

ABSTRACT. The article considers a new approach for small area estimation based on a joint modelling of mean and variances. Model parameters are estimated via expectation–maximization algorithm. The conditional mean squared error is used to evaluate the prediction error. Analytical expressions are obtained for the conditional mean squared error and its estimator. Our approximations are second-order correct, an unwritten standardization in the small area literature. Simulation studies indicate that the proposed method outperforms the existing methods in terms of prediction errors and their estimated values.

Key words: bias correction, conditional mean squared error, EM algorithm, empirical Bayes, hierarchical models, numerical integration, sampling variance, small area estimation

1. Introduction

Small area estimation and the statistical techniques therein have been a topic of great interest to applied and theoretical statistician in recent years. The necessity of reliable small area estimates is felt by many agencies, both public and private, for making useful policy decisions. For example, the small area statistics are being used to monitor socio-economic and public health conditions for different sub-populations defined by age, sex and racial groups over small geographical areas.

It is well known that the direct survey estimates for small areas are usually unreliable, being accompanied with large standard errors and coefficients of variation. Therefore, it is necessary to use models, either explicit or implicit, to connect the small areas, and obtain estimates of improved precision by 'borrowing strength' across areas. The survey-based direct small area estimates and their variance estimates are main ingredients to build area level small area models. Typical modelling strategy assumes that the sampling variances are known, whereas a suitable linear regression model is assumed for the means, and detailed reviews can be found in Ghosh & Rao (1994), Pfeffermann (2002) and Rao (2003). The typical area-level models are subject to two criticisms of interest here: (i) in practice, the sampling variances are estimated quantities, and these are subject to substantial errors because they are often based on equivalent sample sizes as the direct estimates are being calculated; and (ii) the assumption of known and fixed sampling variances does not take into account the uncertainty of (sampling variance) estimation into the overall small area estimation strategy.

The development in small area literature, so far, can be 'loosely' viewed as (i) shrinkage estimation of small area means without variance modelling; (ii) smoothing the direct sampling error variances (not necessarily Bayes or empirical Bayes) to obtain stable variance estimates; and (iii) accounting uncertainty of sampling variance estimation while applying the Fay–Herriot model for mean estimation. Arora & Lahiri (1997), You & Chapman (2006), Liu *et al.* (2007), Dass *et al.* (2012) among others, used sampling variance model in conjunction to small area modelling and hence addressed both the criticisms. Wang & Fuller (2003) provided improved mean squared error (MSE) of small area predictors when direct area variances are estimated. Rivest & Vandal (2003) provided similar results on MSE estimation when the unknown sampling variance is approximated by a normally distributed statistic. Thus, their results addressed the second criticism but not the first one. Otto & Bell (1995), Gershunskaya & Lahiri (2005), Huff *et al.* (2002), Cho *et al.* (2002) and Eltinge *et al.* (2002) explored modelling of direct survey variance estimates. The variance estimation approach considered by Otto & Bell (1995) and Gershunskaya & Lahiri (2005) provides Bayesian or empirical Bayes smoothing of the sampling variances. Thus, their models addressed the first criticism but did not address the second criticism.

For a good discussion on the aforementioned articles and for further references, we refer to Bell (2008). Besides nicely summarizing the latest developments in this aspects of small area estimation, Dr Bell also examined the consequences of this issue in the context of MSE estimation of model-based small area estimators. He further provided numerical evidence of the effect of assuming known sampling variance for the estimation of MSE in the context of Fay–Herriot model (given in (1)).

In our opinion, very less attention has been given for accounting sampling variance estimation effectively while modelling the mean compared with the amount of research devoted to modelling and inferring the small area means. We also feel that there is a lack of systematic development in the small area literature that includes 'shrinking' both means and variances simultaneously. This motivates us to exploit the technique of 'borrowing strength' from other small areas to 'improve' variance estimates as we do for 'improving' the small area mean estimates. We develop a hierarchical model that uses both the point estimates (direct survey-based) and sampling variance estimates to infer all model parameters that determine the stochastic system. Our methodological goal is to develop the dual 'shrinkage' estimation for both the small area means and variances, exploiting the structure of the simultaneous mean-variance modelling so that the final small area estimators are more precise. Numerical evidence shows the effectiveness of dual shrinkage on small area estimates of the mean in terms of prediction error criteria.

Although our modelling perspective is closely related to You & Chapman (2006), their model for true sampling variance should be treated as a prior distribution and hence involve hyperparameters. So one has to use suitable values of the hyperparameters, which could be sensitive from the user's perspective. In our case, this is part of the model and estimated from the likelihood. In their case, another layer of prior distribution is needed if one likes to extend the variance modelling to include regression. But this feature could be easily adopted in our model and estimation. Also they developed Markov Chain Monte Carlo-based full Bayesian approach for model implementation, whereas we developed an empirical Bayes or arguably a frequentist approach where the model parameters are estimated by maximizing the likelihood.

For measuring the uncertainty of the small area estimators, we used the conditional MSE of prediction (CMSEP) to assess the uncertainty of small area estimators. Booth & Hobert (1998) argued strongly for CMSEP as opposed to unconditional MSE to assess the prediction errors for generalized linear mixed models. Recently, the technique has been emphasized by Lohr & Rao (2009) in the context of nonlinear mixed effect models. These authors favoured CMSEP particularly for non-normal models when the conditional variance of the random effects depends on the data. Our model is different from Booth & Hobert (1998) or Lohr & Rao (2009). Hence, the contribution is unique.

A brief outline of the remainder of this article is as follows. In Section 2, we introduce our model. Section 3 describes the method of estimation. In Section 4, we provide an estimation of prediction error in terms of CMSEP. A simulation study has been conducted in Section 5. Section 6 contains some concluding remarks.

2. Model and assumptions

Suppose that there are n small areas, and let $\left(X_i, S_i^2\right)$ be the pair of direct survey estimate and sampling variance for the i-th small area, $i = 1, 2, \cdots, n$. Let $\boldsymbol{Z}_i = (Z_{i1}, \cdots, Z_{ip})^T$ be the vector of p covariates available at the estimation stage. We consider the following hierarchical model:

$$\left.\begin{aligned} X_i \mid \theta_i, \sigma_i^2 &\sim \text{Normal}\left(\theta_i, \sigma_i^2\right) \\ \theta_i &\sim \text{Normal}\left(\boldsymbol{Z}_i^T \boldsymbol{\beta}, \tau^2\right) \end{aligned}\right\}, \tag{1}$$

$$\left.\begin{aligned} \tfrac{(n_i-1)S_i^2}{\sigma_i^2} \mid \sigma_i^2 &\sim \chi^2_{n_i-1} \\ \sigma_i^{-2} &\sim \text{Gamma}(\alpha, \gamma), \end{aligned}\right\}, \tag{2}$$

independently for $i = 1, 2, \cdots, n$. Here, n_i is the sample size for a simple random sample in the i-th area, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_p)^T$ is the $p \times 1$ vector of regression coefficients and $\boldsymbol{B} \equiv \left(\alpha, \gamma, \boldsymbol{\beta}, \tau^2\right)^T$ is the collection of all unknown parameters in the model. Also, Gamma(α, γ) is the Gamma density function with positive shape and scale parameters α and γ, respectively, defined as $f(x) = \{\gamma^\alpha \Gamma(\alpha)\}^{-1} e^{-x/\gamma} x^{\alpha-1}$ for $x > 0$, and 0 otherwise. The unknown σ_i^2 is the true variance of X_i and is usually estimated by the sample variance S_i^2. The model (1), popularly known as Fay–Herriot model, has been widely used in small area literature; see Rao (2003). The modelling of sampling variance in (2) is valid only under simple random sampling with independent and identically distributed normal data in which case the S_i^2's are the sample variances. This assumption is not exactly valid for most of the applications. However, the chi-squared distribution with a careful determination of the degrees of freedom can provide a reasonably useful approximation. Wang & Fuller (2003) suggested the degrees of freedom estimation by approximating the sampling distribution by an 'appropriate' chi-squared distribution. The issue has also been discussed in Maples *et al.* (2009). See remarks 1 and 2 for further discussion on model assumptions. The second level modelling of σ_i^{-2} in (2) can be further extended to $\sigma_i^{-2} \sim$ Gamma($\exp\left(\boldsymbol{Z}_i^T \boldsymbol{\beta}_2\right)/\gamma,\ \gamma)$ (so that $E\left(\sigma_i^{-2}\right) = \exp\left(\boldsymbol{Z}_i^T \boldsymbol{\beta}_2\right)$) for another set of p regression coefficients $\boldsymbol{\beta}_2$ to accommodate the covariate information in the variance modelling.

Although our models are motivated by Hwang *et al.* (2009), we like to mention that Hwang *et al.* (2009) considered shrinking means and variances in the context of microarray data where they prescribed an important solution by plugging in a shrinkage estimator of variance into the mean estimator. The shrinkage estimator of variances in Hwang *et al.* (2009) is a function of S_i^2 only, and not of both X_i and S_i^2. Thus, the shortcomings of their method are as follows: (i) sampling variance estimation is independent of means; (ii) inference on means does not take into account the full uncertainty in variance estimation; and (iii) their model does not include any covariate information.

In our model formulation, the inference for the small area mean parameter θ_i is based on the conditional distribution of θ_i given all the data $\left\{\left(X_i, S_i^2, \boldsymbol{Z}_i\right), i = 1, \cdots, n\right\}$. However, the conditional distribution of θ_i is a non-standard distribution and does not have a closed form expression, thus requiring numerical methods, and for estimation of parameters, we adopt the expectation–maximization (EM) algorithm. The details are provided in the next section.

3. Method of estimation

3.1. *Estimation of the small area parameters*

If the parameters $\boldsymbol{B}$ are known, the joint distribution of $\{X_i, S_i^2, \theta_i, \sigma_i^2\}$ is

$$\begin{aligned}\pi\left(X_i, S_i^2, \theta_i, \sigma_i^2|\boldsymbol{B}\right) &= \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{(X_i-\theta_i)^2}{2\sigma_i^2}\right\} \\ &\times \frac{1}{\Gamma\{(n_i-1)/2\}2^{(n_i-1)/2}}\left[(n_i-1)\frac{S_i^2}{\sigma_i^2}\right]^{(n_i-1)/2-1} \\ &\times \exp\left\{-\frac{(n_i-1)S_i^2}{2\sigma_i^2}\right\}\left(\frac{n_i-1}{\sigma_i^2}\right)\frac{1}{\sqrt{2\pi\tau^2}} \\ &\times \exp\left\{-\frac{(\theta_i-\boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}\right\}\frac{1}{\Gamma(\alpha)\gamma^\alpha}\left(\frac{1}{\sigma_i^2}\right)^{\alpha+1}\exp\left\{-\frac{1}{\gamma\sigma_i^2}\right\} \\ &\propto \exp\left\{-\frac{(X_i-\theta_i)^2}{2\sigma_i^2}-\frac{(n_i-1)S_i^2}{2\sigma_i^2}\right. \\ &\quad \left.-\frac{(\theta_i-\boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}-\frac{1}{\gamma\sigma_i^2}\right\}\left(\frac{1}{\sigma_i^2}\right)^{\frac{n_i}{2}+\alpha+1}\frac{1}{\sqrt{\tau^2}\Gamma(\alpha)\gamma^\alpha} \\ &= \exp\left[-\frac{(\theta_i-\boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau^2}-\left\{\frac{(X_i-\theta_i)^2}{2}\right.\right. \\ &\quad \left.\left.+\frac{(n_i-1)S_i^2}{2}+\frac{1}{\gamma}\right\}\frac{1}{\sigma_i^2}\right]\left(\frac{1}{\sigma_i^2}\right)^{\frac{n_i}{2}+\alpha+1}\frac{1}{\sqrt{\tau^2}\Gamma(\alpha)\gamma^\alpha}.\end{aligned} \tag{3}$$

Therefore, the conditional distribution of σ_i^2 and θ_i given the data (X_i, S_i^2), $i = 1, \cdots, n$ and $\boldsymbol{B}$ are

$$\pi\left(\sigma_i^2|X_i, S_i^2, \boldsymbol{B}\right) \propto \frac{\exp\left[-\frac{\left(X_i-\boldsymbol{Z}_i^T\boldsymbol{\beta}\right)^2}{2(\sigma_i^2+\tau^2)}-\left\{\frac{(n_i-1)}{2}S_i^2+\frac{1}{\gamma}\right\}\frac{1}{\sigma_i^2}\right]}{\left(\sigma_i^2\right)^{(n_i-1)/2+\alpha+1}\left(\sigma_i^2+\tau^2\right)^{1/2}}, \tag{4}$$

$$\pi\left(\theta_i|X_i, S_i^2, \boldsymbol{B}\right) \propto \exp\left\{-\frac{\left(\theta_i-\boldsymbol{Z}_i^T\boldsymbol{\beta}\right)^2}{2\tau^2}\right\}\psi_i^{-(n_i/2+\alpha)}, \tag{5}$$

where $\psi_i = 0.5(X_i-\theta_i)^2 + 0.5(n_i-1)S_i^2 + 1/\gamma$. Note that the aforementioned conditionals are obtained by integrating out θ_i and σ_i^2, respectively, from the joint distribution (3). Note that (4) can be used to estimate the true sampling variances. The estimator is clearly dependent on both the mean and sampling variance estimates.

From now on, we will borrow the notations from Booth & Hobert (1998) for determining all related stochastic distributions. In this context, a meaningful point estimator for θ_i is its conditional mean,

$$\theta_i\left(\boldsymbol{B}; X_i, S_i^2\right) = E_{\boldsymbol{B}}\left(\theta_i|X_i, S_i^2\right), \tag{6}$$

where $E_{\boldsymbol{B}}$ represents the expectation with respect to the conditional distribution of θ_i with known $\boldsymbol{B}$. A sensible prediction error can be measured by the posterior variance

$$\nu_i\left(\boldsymbol{B}; X_i, S_i^2\right) = \mathrm{var}_{\boldsymbol{B}}\left(\theta_i | X_i, S_i^2\right). \tag{7}$$

Neither (6) nor (7) has any closed form expression. Therefore, we have used numerical quadrature to compute these integrals. In practice, $\boldsymbol{B}$ is unknown. We estimate them by maximizing the marginal likelihood, and the details are given in the next section. Let $\hat{\boldsymbol{B}}$ denote the corresponding estimator. Substituting $\boldsymbol{B}$ by $\hat{\boldsymbol{B}}$ in formulas (6) and (7) will produce the estimates of θ_i and ν_i:

$$\hat{\theta}_i = \theta_i(\hat{\boldsymbol{B}}; X_i, S_i^2) = E_{\hat{\boldsymbol{B}}}(\theta_i | X_i, S_i^2), \tag{8}$$

$$\hat{\nu}_i = \nu_i(\hat{\boldsymbol{B}}; X_i, S_i^2) = \mathrm{var}_{\hat{\boldsymbol{B}}}(\theta_i | X_i, S_i^2). \tag{9}$$

The estimator (8) is popularly known as the empirical Bayes estimator, and (9) is the estimated Bayes risks. It is well known that the $\hat{\nu}_i$ only considered the variability in the prediction procedure, but not the variability due to the parameter estimation ($\hat{\boldsymbol{B}}$). We accounted this additional variability by adapting the technique of Booth & Hobert (1998) in this set-up. The details are discussed in Section 4.1.

3.2. *Estimation of the structural parameters*

We obtain the maximum likelihood estimate of $\boldsymbol{B}$ by maximizing the marginal likelihood $L_M = \prod_{i=1}^n L_i^M$ of $\left\{\left(X_i, S_i^2, \boldsymbol{Z}_i\right)_{i=1}^n; \boldsymbol{B}\right\}$, where

$$L_i^M \propto \frac{\Gamma\left(\frac{n_i}{2}+\alpha\right)}{\sqrt{\tau^2}\Gamma(\alpha)\gamma^\alpha} \int \exp\left\{-\frac{\left(\theta_i - \boldsymbol{Z}_i^T \boldsymbol{\beta}\right)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+\alpha\right)} d\theta_i.$$

Note that L_M is eventually a function of $\alpha, \gamma, \boldsymbol{\beta}$ and τ^2. Direct maximization of L_M is difficult as it involves non-standard integrals with respect to unobserved θ_is. Therefore, we adopt the EM algorithm to estimate the parameters iteratively. The EM algorithm consists of two steps, E and M steps, and the unobserved θ_is are treated as missing data. Note that L_i^M is the observed data likelihood, and for the EM algorithm, we also need complete data likelihood $L_{i,\mathrm{compl}}$ in terms of complete data $\theta_i, X_i, \boldsymbol{Z}_i$, where

$$L_{i,\mathrm{compl}} \propto \frac{\Gamma\left(\frac{n_i}{2}+\alpha\right)}{\sqrt{\tau^2}\Gamma(\alpha)\gamma^\alpha} \exp\left\{-\frac{\left(\theta_i - \boldsymbol{Z}_i^T \boldsymbol{\beta}\right)^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+\alpha\right)}.$$

In the E step of the t^{th} iteration, we take expectation of the complete data log likelihood with respect to the conditional density of the unobserved variable, which is θ_i in our case. In the M step, we maximize the expected log complete data likelihood with respect to the unknown parameters keeping the conditional expectations fixed. We repeat these two steps until the parameter estimates converge. Let $\hat{\boldsymbol{\beta}}^{(t)}$, $\left(\hat{\tau}^2\right)^{(t)}$, $\hat{\alpha}^{(t)}$ and $\gamma^{(t)}$ be the estimates of $\boldsymbol{\beta}, \alpha$ and γ, respectively, at the t^{th} iteration. Then, at the t^{th} iteration of the EM algorithm,

$$\hat{\boldsymbol{\beta}}^{(t)} = \left(\sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^T\right)^{-1} \left\{\sum_{i=1}^n \boldsymbol{Z}_i E^{(t-1)}(\theta_i)\right\}, (\hat{\tau}^2)^{(t)} = \sum_{i=1}^n E^{(t-1)}\left(\theta_i - \boldsymbol{Z}_i^T \boldsymbol{\beta}\right)^2 / n,$$

and $\hat{\alpha}^{(t)}$ and $\hat{\gamma}^{(t)}$ are estimated by solving $S_\alpha = 0$ and $S_\gamma = 0$ where

$$S_\alpha = \sum\nolimits_{i=1}^{n} \left[\partial \log \left\{ \Gamma(n_i/2 + \alpha) \right\} / \partial\alpha - \partial \log \left\{ \Gamma(\alpha) \right\} / \partial\alpha - \log \gamma - E^{(t-1)} \left\{ \log(\psi_i) \right\} \right],$$

$$S_\gamma = -n\alpha/\gamma + \sum_{i=1}^{n} \left\{ (n_i/2) + \alpha \right\} E^{(t-1)} \left(\psi_i^{-1} \right) / \gamma^2.$$

The expectations at the t^{th} iteration, $E^{(t-1)}$, are the expectations with respect to the conditional density of θ_i, $\pi \left(\theta_i | X_i, S_i^2, \boldsymbol{B}^{(t-1)} \right)$, $i = 1, \ldots, n$, with $\boldsymbol{B}^{(t-1)}$ denoting the estimate of $\boldsymbol{B}$ at the $(t-1)^{th}$ iteration. Because the conditional density is a non-standard density, one may use the Monte Carlo method (such as rejection sampling) or numerical integration to evaluate those integrals. We used Gauss–Hermite 20-point numerical integration. The details on how we obtain S_α and S_γ are given in the Supporting information (Supplementary Appendix).

4. Prediction error calculation

4.1. *Mean squared error of prediction*

Following the definition of CMSEP of Booth & Hobert (1998), the prediction error variance is

$$\begin{aligned} CMSEP \left(\boldsymbol{B}; X_i, S_i^2 \right) &= E_{\boldsymbol{B}} \left[\left(\hat{\theta}_i - \theta_i \right)^2 | X_i, S_i^2 \right] \\ &= E_{\boldsymbol{B}} \left[\left\{ \hat{\theta}_i - \theta_i \left(\boldsymbol{B}, X_i, S_i^2 \right) + \theta_i \left(\boldsymbol{B}, X_i, S_i^2 \right) - \theta_i \right\}^2 | X_i, S_i^2 \right], \end{aligned}$$

where $\hat{\theta}_i$ and $\theta_i \left(\boldsymbol{B}, X_i, S_i^2 \right)$ are defined in (8) and (6), respectively. Because $\theta_i (\boldsymbol{B}; X_i, S_i^2) - \theta_i$ and $\hat{\theta}_i - \theta_i (\boldsymbol{B}; X_i, S_i^2)$ are conditionally uncorrelated given X_i and S_i^2,

$$\begin{aligned} CMSEP \left(\boldsymbol{B}; X_i, S_i^2 \right) &= \mathrm{var}_{\boldsymbol{B}} \left(\theta_i | X_i, S_i^2 \right) + E_{\boldsymbol{B}} \left[\left\{ \hat{\theta}_i - \theta_i \left(\boldsymbol{B}; X_i, S_i^2 \right) \right\}^2 | X_i, S_i^2 \right] \\ &= \nu_i \left(\boldsymbol{B}; X_i, S_i^2 \right) + c_i \left(\boldsymbol{B}; X_i, S_i^2 \right), \end{aligned} \tag{10}$$

where $c_i \left(\boldsymbol{B}; X_i, S_i^2 \right)$ is the correction term due to the estimation of unknown parameters $\boldsymbol{B}$. The correction contribution is of order $O_p(n^{-1})$. Note that the aforementioned measure is still not usable because it involves the unknown structural parameters $\boldsymbol{B}$. It is natural to plug-in the estimate $\hat{\boldsymbol{B}}$ of $\boldsymbol{B}$ and get an usable measure of mean squared prediction error

$$\widehat{CMSEP}_i = MSE \left(\hat{\boldsymbol{B}}; X_i, S_i^2 \right) = \nu_i \left(\hat{\boldsymbol{B}}; X_i, S_i^2 \right) + c_i \left(\hat{\boldsymbol{B}}; X_i, S_i^2 \right) = \hat{\nu}_i + \hat{c}_i.$$

As it will be clear in the next section (as well as from small area estimation literature), the estimator (10) has considerable bias. Typically, the order of the bias is $O_p(n^{-1})$ due to estimation of ν_i by $\hat{\nu}_i$.

4.2. Bias correction for $\nu_i\left(\hat{\boldsymbol{B}}; X_i, S_i^2\right)$

For the bias correction of the estimated conditional variance $\nu_i\left(\hat{\boldsymbol{B}}; X_i, S_i^2\right)$, we expand this about $\boldsymbol{B}$

$$\begin{aligned}\hat{\nu}_i &= \nu_i\left(\hat{\boldsymbol{B}}; X_i, S_i^2\right)\\ &= \nu_i\left(\boldsymbol{B}; X_i, S_i^2\right) + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T \frac{\partial \nu_i\left(B; X_i, S_i^2\right)}{\partial \boldsymbol{B}}\\ &\quad + \frac{1}{2}(\hat{\boldsymbol{B}} - \boldsymbol{B})^T \frac{\partial \nu_i\left(\boldsymbol{B}; X_i, S_i^2\right)}{\partial \boldsymbol{B} \partial \boldsymbol{B}^T}(\hat{\boldsymbol{B}} - \boldsymbol{B}) + O_p(n^{-2}).\end{aligned} \tag{11}$$

Then, the approximated bias involved in $\nu_i\left(\hat{\boldsymbol{B}}; X_i, S_i^2\right)$ is

$$E(\hat{\boldsymbol{B}} - \boldsymbol{B}) \frac{\partial \nu_i\left(\boldsymbol{B}; X_i, S_i^2\right)}{\partial \boldsymbol{B}} + \frac{1}{2}\mathrm{tr}\left\{\frac{\partial \nu_i\left(\boldsymbol{B}; X_i, S_i^2\right)}{\partial \boldsymbol{B} \partial \boldsymbol{B}^T} \boldsymbol{I}^{-1}(\boldsymbol{B})\right\},$$

where $\boldsymbol{I}(\boldsymbol{B})$ is the Fisher's information matrix obtained from the marginal likelihood L_M given in the beginning of Section 3.2.

Handling the bias analytically is difficult. Booth & Hobert (1998) adopted bootstrap bias correction approach in a similar scenario. This is because there is no closed form expressions available for this bias terms. The bootstrap method requires repeated estimation of model parameters on the basis of a re-sampled data. This often pauses practical and computational difficulties in this hierarchical model. As we will see in the next subsection, handling the second term in (10) is also difficult because of the same difficulty of not having any closed form expression. Thus, if we have to do the bootstrap for the bias correction of ν_i, the estimation of c_i can also be performed in the same run. It is not necessary to obtain any analytical approximations. The resampling techniques has been used in Jiang *et al.* (2002) and Hall & Maiti (2006). Here, we derived Taylor expansion-based approximations and handle the inside integrals by applying the Gauss–Hermite quadrature formula and thereby avoiding the repeated evaluation of model parameters. The estimation becomes computationally fast.

Let $\hat{\boldsymbol{B}}$ be the maximum likelihood estimator of $\boldsymbol{B}$ as proposed in the previous section. Following Cox & Snell (1968, Equation (20)), we approximate $E\left(\hat{\boldsymbol{B}} - \boldsymbol{B}\right)$ up to $O(n^{-1})$. Taking $\boldsymbol{I} \equiv \boldsymbol{I}(\boldsymbol{\beta})$, define $\boldsymbol{I}^{-1} = ((I^{rs}))$ as the inverse of $\boldsymbol{I} = ((I_{rs}))$, where $I_{rs} = E\left(-V_{rs}^{(1)}\right)$ and $V_{rs}^{(i)} = \partial^2 \log L_i^M / \partial \boldsymbol{B}_r \partial \boldsymbol{B}_s$. The bias in the s^{th} element of $\hat{\boldsymbol{B}}$ is

$$E\left(\hat{\boldsymbol{B}}_s - \boldsymbol{B}_s\right) \approx \frac{1}{2}\sum_r \sum_t \sum_u I^{rs} I^{tu}(K_{rtu} + 2J_{t,ru}), \tag{12}$$

$$K_{rst} = E\left(W_{rst}^{(\cdot)}\right), W_{rst}^{(i)} = \frac{\partial^3 \log L_i^M}{\partial \boldsymbol{B}_r \partial \boldsymbol{B}_s \partial \boldsymbol{B}_t}, J_{r,st} = E\left\{\sum U_r^{(i)} V_{st}^{(i)}\right\}, U_r^{(i)} = \frac{\partial \log L_i^M}{\partial \boldsymbol{B}_r},$$

where the L_i^M is the marginal likelihood of $\left(X_i, S_i^2\right)$ defined in the previous section. The detailed formulas are given in the Supporting information (Supplementary Appendix).

With

$$\left(\frac{\partial \nu_i}{\partial \boldsymbol{B}}\right)^T = \left(\frac{\partial \nu_i}{\partial \alpha}, \frac{\partial \nu_i}{\partial \gamma}, \frac{\partial \nu_i}{\partial \boldsymbol{\beta}}, \frac{\partial \nu_i}{\partial \tau^2}\right),$$

and $\nu_i\left(\boldsymbol{B};X_i,S_i^2\right)=\mathrm{var}_{\boldsymbol{B}}\left(\theta_i|X_i,S_i^2\right)$, we have

$$\begin{aligned}
\frac{\partial\nu_i}{\partial\alpha}&=-\mathrm{cov}^*\left\{\theta_i^2,\log(\psi_i)\right\}+2E^*(\theta_i)\mathrm{cov}^*\{\theta_i,\log(\psi_i)\}\\
\frac{\partial\nu_i}{\partial\gamma}&=\left(\frac{n_i}{2}+\alpha\right)\frac{1}{\gamma^2}\left\{\mathrm{cov}^*\left(\theta_i^2,\frac{1}{\psi_i}\right)-2E^*(\theta_i)\mathrm{cov}^*\left(\theta_i,\frac{1}{\psi_i}\right)\right\}\\
\frac{\partial\nu_i}{\partial\boldsymbol{\beta}}&=\frac{1}{\tau^2}\left\{\mathrm{cov}^*\left(\theta_i^2,\theta_i\right)-2E^*(\theta_i)\mathrm{cov}^*(\theta_i,\theta_i)\right\}\\
\frac{\partial\nu_i}{\partial\tau^2}&=\frac{1}{2(\tau^2)^2}\left[\mathrm{cov}^*\left\{\theta_i^2,\left(\theta_i-\boldsymbol{Z}_i^T\boldsymbol{\beta}\right)^2\right\}-2E^*(\theta_i)\mathrm{cov}^*\left\{\theta_i,\left(\theta_i-\boldsymbol{Z}_i^T\boldsymbol{\beta}\right)^2\right\}\right],
\end{aligned}$$

where the * means that the expectation, variance and covariance are calculated with respect to the conditional distribution of θ_i at the estimated parameters' value. The approximated expression of $\partial\nu_i\left(\boldsymbol{B};X_i,S_i^2\right)/\partial\boldsymbol{B}\partial\boldsymbol{B}^T$ is given in the Supporting information. The expectation, variance and covariance are computed on the basis of the numerical integration. Therefore,

$$\hat{\nu}_i\approx\nu_i\left(\hat{\boldsymbol{B}};X_i,S_i^2\right)-(\hat{\boldsymbol{B}}-\boldsymbol{B})\frac{\partial\nu_i}{\partial\boldsymbol{B}}-\frac{1}{2}\mathrm{tr}\left\{\frac{\partial\nu_i\left(\boldsymbol{B};X_i,S_i^2\right)}{\partial\boldsymbol{B}\partial\boldsymbol{B}^T}\boldsymbol{I}^{-1}(\boldsymbol{B})\right\}.$$

This expression is second-order correct meaning that the bias of this term is of order $o_p(n^{-1})$.

4.3. *Approximation of* $c_i\left(\boldsymbol{B};X_i,S_i^2\right)$

The definition of $c_i\left(\boldsymbol{B};X_i,S_i^2\right)$ is given in the previous section,

$$c_i\left(\boldsymbol{B};X_i,S_i^2\right)=E_{\boldsymbol{B}}\left[\left\{\hat{\theta}_i-\theta_i\left(\boldsymbol{B};X_i,S_i^2\right)\right\}^2|X_i,S_i^2\right],$$

where $\hat{\theta}_i-\theta_i\left(\boldsymbol{B};X_i,S_i^2\right)=\theta_i\left(\hat{\boldsymbol{B}};X_i,S_i^2\right)-\theta_i\left(\boldsymbol{B};X_i,S_i^2\right)$. Using the Taylor series expansion and ignoring the term of the order $O_p(|\hat{\boldsymbol{B}}-\boldsymbol{B}|^2)$, we write

$$\theta_i\left(\hat{\boldsymbol{B}};X_i,S_i^2\right)-\theta_i\left(\boldsymbol{B};X_i,S_i^2\right)\approx A_i^T\left(\boldsymbol{B};X_i,S_i^2\right)(\hat{\boldsymbol{B}}-\boldsymbol{B}),\tag{13}$$

where

$$\begin{aligned}
A_i\left(\boldsymbol{B};X_i,S_i^2\right)^t&=\frac{\partial\theta_i\left(\boldsymbol{B};X_i,S_i^2\right)}{\partial\boldsymbol{B}}\\
&=\left(\frac{\partial\theta_i\left(\boldsymbol{B};X_i,S_i^2\right)}{\partial\alpha},\frac{\partial\theta_i\left(\boldsymbol{B};X_i,S_i^2\right)}{\partial\gamma},\frac{\partial\theta_i\left(\boldsymbol{B};X_i,S_i^2\right)}{\partial\boldsymbol{\beta}},\frac{\partial\theta_i\left(\boldsymbol{B};X_i,S_i^2\right)}{\partial\tau^2}\right).
\end{aligned}$$

Because $\theta_i\left(\boldsymbol{B}; X_i, S_i^2\right) = E_{\boldsymbol{B}}\left(\theta_i | X_i, S_i^2\right)$, the components of A_i are

$$
\begin{aligned}
\frac{\partial \theta_i\left(\boldsymbol{B}; X_i, S_i^2\right)}{\partial \alpha} &= E^*(\theta_i)E^*\left\{\log(\psi_i)\right\} - E^*\left\{\theta_i \log(\psi_i)\right\} = -\text{cov}^*\left\{\theta_i, \log(\psi_i)\right\}, \\
\frac{\partial \theta_i\left(\boldsymbol{B}; X_i, S_i^2\right)}{\partial \gamma} &= \left(\frac{n_i}{2} + \alpha\right)\frac{1}{\gamma^2}\left\{E^*\left(\frac{\theta_i}{\psi_i}\right) - E^*(\theta_i)E^*\left(\frac{1}{\psi_i}\right)\right\} \\
&= \left(\frac{n_i}{2} + \alpha\right)\frac{1}{\gamma^2}\text{cov}^*\left(\theta_i, \frac{1}{\psi_i}\right), \\
\frac{\partial \theta_i\left(\boldsymbol{B}; X_i, S_i^2\right)}{\partial \boldsymbol{\beta}} &= \frac{1}{\tau^2}\left[E^*\left(\theta_i^2\right) - \{E^*(\theta_i)\}^2\right] = \frac{1}{\tau^2}\text{var}^*(\theta_i), \\
\frac{\partial \theta_i\left(\boldsymbol{B}; X_i, S_i^2\right)}{\partial \tau^2} &= \frac{1}{2(\tau^2)^2}\left[E^*\left\{\theta_i\left(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}\right)^2\right\} - E^*(\theta_i)E^*\left(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}\right)^2\right] \\
&= \frac{1}{2\tau^2}\text{cov}^*\left\{\theta_i, \left(\theta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}\right)^2\right\}.
\end{aligned}
$$

As a consequence of (13),

$$
c_i\left(\boldsymbol{B}; X_i, S_i^2\right) \approx A_i\left(\boldsymbol{B}; X_i, S_i^2\right)^T I(\boldsymbol{B})^{-1} A_i\left(\boldsymbol{B}; X_i, S_i^2\right), \tag{14}
$$

and the approximation is correct up to the order $O_p(n^{-1})$. Substituting $\hat{\boldsymbol{B}}$ into formula (14) will yield an estimate of the correction term,

$$
c_i\left(\hat{\boldsymbol{B}}; X_i, S_i^2\right) \approx A_i\left(\hat{\boldsymbol{B}}; X_i, S_i^2\right)^T I(\hat{\boldsymbol{B}})^{-1} A_i\left(\hat{\boldsymbol{B}}; X_i, S_i^2\right).
$$

As the estimated information matrix is $\left\{I\left(\hat{\boldsymbol{B}}\right)\right\}^{-1} = O_p(n^{-1})$, the error in the approximation is of order $o_p(n^{-1})$. Similar to the case of ν_i, the entries in $A_i\left(\hat{\boldsymbol{B}}; X_i, S_i^2\right)$ do not have closed forms. They are approximated by numerical integrals. By summing up all the derivations and approximations, we obtain the following result.

Theorem 1. *The estimated CMSEP for $\hat{\theta}_i$ is*

$$
\begin{aligned}
\widehat{CMSEP} = \nu_i\left(\hat{\boldsymbol{B}}; X_i, S_i^2\right) - (\hat{\boldsymbol{B}} - \boldsymbol{B})\frac{\partial \nu_i}{\partial \boldsymbol{B}}|_{B=\hat{B}} - \frac{1}{2}\text{tr}\left\{\frac{\partial \nu_i\left(\boldsymbol{B}; X_i, S_i^2\right)}{\partial \boldsymbol{B}\partial \boldsymbol{B}^T}\boldsymbol{I}^{-1}(\boldsymbol{B})\right\}|_{B=\hat{B}} \\
+ A_i\left(\hat{\boldsymbol{B}}; X_i, S_i^2\right)^T I(\hat{\boldsymbol{B}})^{-1} A_i\left(\hat{\boldsymbol{B}}; X_i, S_i^2\right).
\end{aligned}
$$

The error is $o_p(n^{-1})$.

Remark 1. As mentioned in Section 2, the degree of freedom associated with the χ^2 distribution for the sampling variance is not simply $n_i - 1$, n_i being the sample size for i-th area, except the scenario when the survey design is a simple random sampling. There is no sound theoretical result for determining the degree of freedom when the survey design is complex. Wang & Fuller (2003) suggested a moment-based estimation of the degrees of freedom that requires the fourth moment of sampling error distributions. For example, Equation (44) in their paper, suggested a degree-of-freedom estimator, $\hat{d}_i = \hat{\kappa}_{4i}^{-1}2(n_i - 1)$ where $\hat{\kappa}_{4i}$ is an estimator of the standardized fourth moment of the error distribution. Thus, in this case, the approximated distribution of S_i^2 is $\hat{d}_i^{-1}\chi^2_{\hat{d}_i}\sigma_i^2$. Alternatively, if one knows the exact sampling design, then the simulation-based guideline of Maples *et al.* (2009) could be useful. For county level estimation using the American Community Survey, they suggested the estimated degrees of freedom to be $0.36 \times \sqrt{n_i}$.

The normal distribution-based approximation may not work well unless the sample sizes n_i's are large.

Remark 2. Note that without any hierarchical modelling assumption, S_i^2 and X_i are independent as S_i^2 and X_i are, respectively, ancillary and the complete sufficient statistics for θ_i under the normal distribution assumption. However, under models (1) and (2), the conditional distribution of σ_i^2 and θ_i involves both X_i and S_i^2, which is seen from (4) and (5). Therefore, the shrinkage estimators for both σ_i^2 and θ_i naturally involve all the observed data. Furthermore, the estimated structural parameters are also functions of X_i and S_i^2. Thus, our estimation is mean-variance integrated as opposed to augmented.

5. Simulation study

Simulation design : In order to study finite sample performance of the proposed estimators, a simulation set-up closely related to Wang & Fuller (2003) was considered. To simplify the simulations, we set $Z = 1$ and did not choose any other covariate; only $\left(X_i, S_i^2\right)$ were generated. First, we generated observations for each small area using the model

$$X_{ij} = \beta + u_i + e_{ij}, j = 1, \dots, n_i,\ i = 1, \dots, n,$$

where $u_i \sim \text{Normal}(0, \tau^2)$ and $e_{ij} \sim \text{Normal}\left(0, n_i\sigma_i^2\right)$. Then the random effects model for the small area mean is

$$X_i = \beta + u_i + e_i,\quad i = 1, \dots, n,$$

where $X_i = \bar{X_i}. = n_i^{-1}\sum_{j=1}^{n_i} X_{ij}, e_i = \bar{e_i}. = n_i^{-1}\sum_{j=1}^{n_i} e_{ij}$. Therefore, $X_i \sim \text{Normal}(\theta_i, \sigma_i^2)$, where $\theta_i = \beta + u_i$; $\theta_i \sim \text{Normal}(\beta, \tau^2)$, and $e_i \sim \text{Normal}\left(0, \sigma_i^2\right)$. The parameter of interest is the mean, θ_i, for the i-th small area, $i = 1, \dots, n$. The direct estimator for σ_i^2 we used is

$$S_i^2 = \frac{1}{n_i - 1}\frac{1}{n_i}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2.$$

It is to be noted that $(n_i - 1)S_i^2/\sigma_i^2 \sim \chi^2_{(n_i-1)}$. Like Wang & Fuller (2003), we set all n_i equal to m that eased our programming efforts. However, the sampling variances were still unequal by choosing one-third of the σ_i^2 that was equal to 1, one-third was equal to 4 and the rest were equal to 16. In the simulations, we set $\beta = 10$ and took three different values of τ^2 as 0.5, 1 and 4. For each of the τ^2, we generated $N = 10\,000$ samples for each of the combinations $(m, n) = (9, 36)$, $(18, 180)$, $(40, 36)$ and $(40, 180)$. In Table 1, we present the mean and standard deviation over the simulations of the estimates of β and τ^2 on the basis of the proposed method. The maximum likelihood estimators of the model parameters perform reasonably well, even for small samples. It is known that the τ^2 is difficult to estimate from a small sample when its true value is small. In our case, this estimate is, in fact, very reasonable. One reason could be, although these are maximum likelihood estimators, unlike the standard mean model (Fay–Herriot), because we used the joint distribution of $\left(X_i, S_i^2\right)$.

Next, we report the performance on point prediction and MSE of prediction (MSEP) estimation. The results are averaged over areas within the group having same sampling variances. We will denote our method as method I, whereas the point estimators and the MSEP estimators obtained from Wang & Fuller (2003) are referred to as method II. Because we do not have the access of Wang and Fuller's computing code, we reproduced all the numerical results using a code written by us. For the sake of clarity, we copied their reported values whenever available. They are in Tables 2–5 within parentheses.

Table 1. *Results of the simulation study*

τ^2	Mean	SD	Mean	SD
	$n = 36, m = 9$		$n = 180, m = 18$	
		For β		
0.5	9.994	0.404	10.000	0.185
1	10.004	0.399	10.004	0.185
4	9.999	0.490	9.996	0.220
		For τ^2		
0.5	0.536	0.050	0.522	0.015
1	1.087	0.144	1.054	0.043
4	4.299	0.853	4.230	0.269
	$n = 36, m = 40$		$n = 180, m = 40$	
		For β		
0.5	9.993	0.407	10.000	0.185
1	9.996	0.404	10.002	0.186
4	9.996	0.493	9.998	0.220
		For τ^2		
0.5	0.515	0.032	0.511	0.013
1	1.033	0.091	1.027	0.038
4	4.084	0.559	4.105	0.256

Here, we present average and standard deviation (SD) on the basis of simulation of the estimates of β and τ^2. We set $\beta = 10$.

Table 2. *Results of the simulation study when* $n = 36$ *and* $m = 9$

		$\tau^2 = 0.5$		$\tau^2 = 1$		$\tau^2 = 4$	
	σ_i^2	I	II	I	II	I	II
			Performance of the point predictor				
Bias	1	−0.004	−0.005	0.002	0.001	−0.001	−0.001
	4	−0.005	−0.003	0.003	0.002	0.007	0.007
	16	−0.007	−0.006	0.002	−0.001	−0.001	−0.002
MSEP	1	0.422	0.491 (0.493)	0.564	0.623 (0.633)	0.846	0.859 (0.859)
	4	0.587	0.770 (0.755)	0.937	1.131 (1.110)	2.179	2.293 (2.290)
	16	0.648	0.841 (0.820)	1.101	1.359 (1.310)	3.535	3.853 (3.760)
			Performance of estimated MSEP				
RB	1	0.028	0.204 (0.156)	0.124	0.085 (0.038)	0.209	0.022 (0.019)
	4	−0.027	0.315 (0.223)	0.040	0.138 (0.072)	0.055	0.009 (−0.026)
	16	−0.031	0.471 (0.329)	0.052	0.232 (0.137)	0.038	0.010 (−0.037)
RRMSEP	1	0.596	0.600	0.309	0.484	0.501	0.446
	4	0.125	0.861	0.178	0.606	0.306	0.372
	16	0.095	1.181	0.140	0.795	0.216	0.378
RSD	1	0.595	0.564	0.283	0.476	0.455	0.446
	4	0.122	0.801	0.173	0.590	0.301	0.372
	16	0.090	1.082	0.130	0.760	0.212	0.378

Note: The values within parentheses are from Wang and Fuller (2003).
The upper panel presents the bias of $\hat{\theta}_i$ and the MSEP of $\hat{\theta}_i$, whereas the lower panel represents the relative bias, the relative root mean squared error and the relative standard deviation of the estimated MSEP. The proposed method and the method of Wang and Fuller are abbreviated as I and II, respectively.
RB, relative bias; MSEP, mean squared error of prediction; RRMSEP, relative root mean squared error of prediction; RDS, relative standard deviation.

Statistics for comparing the point estimators : We report empirical bias and MSEP for comparing the point prediction in Tables 2–5 for various combination of sample sizes. Here, the MSEP for area i is calculated as the average over the 10 000 replications of $\left(\hat{\theta}_i - \theta_i\right)^2$. Although, for large model variance τ^2, both methods provide comparable biases, method I tends to have larger bias for small τ^2. As the sample size increases, bias under both methods reduces.

The most striking outcome of our method is in reducing the MSEP. The MSEP under method I is almost always lower than the MSEP under method II. The ratio of MSEP for method II over MSEP for method I indicates gain in all cases except $\left(m, n, \tau^2, \sigma_i^2\right) = (40, 36, 0.5, 16)$ and $\left(m, n, \tau^2, \sigma_i^2\right) = (40, 180, 0.5, 16)$, where the ratios are 0.974 and 0.993 respectively. In these cases, the ratio of sampling variance to model variance is very high. However, the ratio is 1.313 and 1.085, respectively, for $\left(m, n, \tau^2, \sigma_i^2\right) = (9, 36, 0.5, 4)$ and (18,180, 0.5,4). Thus, there is a gain for small within sample sizes. The maximum reduction in MSEP by method I compared with method II is 31%. The relative bias (RB) (ratio of square of bias to MSEP) is negligible because of low bias in every situation.

Statistics for comparing estimated MSEP : The lower part of Tables 2–5 contains results for comparing estimated MSEP. We used empirical measures of RB and relative root MSE to quantify the performance of MSEP estimators. RB of the MSEP estimator was defined by

$$RB_i = \frac{E\left\{\widehat{MSEP}_i\right\} - MSEP_i}{MSEP_i},$$

Table 3. *Results of the simulation study when* $n = 180$ *and* $m = 18$

		$\tau^2 = 0.5$		$\tau^2 = 1$		$\tau^2 = 4$	
	σ_i^2	I	II	I	II	I	II
			Performance of the point predictor				
Bias	1	0.001	0.001	0.002	0.001	0.000	0.000
	4	0.001	0.001	0.003	0.002	−0.002	−0.002
	16	0.000	−0.001	0.003	0.002	−0.004	−0.003
MSEP	1	0.356	0.390(0.387)	0.521	0.544(0.543)	0.816	0.819(0.818)
	4	0.476	0.516(0.511)	0.838	0.883(0.877)	2.065	2.094(2.100)
	16	0.520	0.532(0.527)	0.978	1.003(1.000)	3.293	3.350(3.350)
			Performance of estimated MSEP				
RB	1	0.027	0.204(0.199)	0.064	0.096(0.087)	0.127	0.015(0.013)
	4	0.007	0.342(0.315)	0.024	0.163(0.140)	0.035	0.026(0.010)
	16	0.007	0.460(0.417)	0.036	0.235(0.200)	0.032	0.044(0.024)
RRMSEP	1	0.118	0.432	0.182	0.297	0.309	0.291
	4	0.052	0.644	0.089	0.370	0.186	0.217
	16	0.031	0.814	0.059	0.475	0.102	0.191
RSD	1	0.115	0.381	0.171	0.281	0.281	0.291
	4	0.052	0.545	0.086	0.333	0.183	0.215
	16	0.030	0.672	0.047	0.413	0.096	0.186

Note: The values within parentheses are from Wang and Fuller (2003).
The upper part presents the bias of $\hat{\theta}_i$ and the MSEP of $\hat{\theta}_i$, where as the lower part represents the relative bias, the relative root mean squared error and the relative standard deviation of estimated MSEP. The proposed method and the method of Wang and Fuller are abbreviated as I and II, respectively.
RB, relative bias; MSEP, mean squared error of prediction; RRMSEP, relative root mean squared error of prediction; RDS, relative standard deviation.

Table 4. *Results of the simulation study when* $n = 36$ *and* $m = 40$

		$\tau^2 = 0.5$		$\tau^2 = 1$		$\tau^2 = 4$	
	σ_i^2	I	II	I	II	I	II
		Performance of the point predictor					
Bias	1	−0.004	−0.002	−0.001	0.000	−0.005	−0.005
	4	−0.008	−0.005	−0.001	−0.001	−0.004	−0.004
	16	−0.007	−0.004	−0.005	−0.004	−0.003	−0.004
MSEP	1	0.414	0.435	0.544	0.589	0.819	0.849
	4	0.579	0.610	0.916	0.984	2.107	2.197
	16	0.643	0.627	1.089	1.120	3.401	3.524
		Performance of estimated MSEP					
RB	1	−0.036	0.197	0.050	0.043	0.064	−0.001
	4	−0.070	0.273	0.008	0.068	0.033	−0.014
	16	−0.081	0.399	0.003	0.137	0.033	−0.004
RRMSEP	1	0.098	0.467	0.142	0.325	0.204	0.207
	4	0.087	0.830	0.082	0.566	0.144	0.256
	16	0.093	1.149	0.073	0.776	0.119	0.365
RSD	1	0.091	0.424	0.133	0.322	0.193	0.207
	4	0.052	0.783	0.081	0.562	0.140	0.256
	16	0.047	1.078	0.073	0.764	0.115	0.365

The upper part presents the bias of $\hat{\theta}_i$ and the MSEP of $\hat{\theta}_i$, where as the lower part represents the relative bias, the relative root mean squared error and the relative standard deviation of estimated MSEP. The proposed method and the method of Wang and Fuller are abbreviated as I and II, respectively.
RB, relative bias; MSEP, mean squared error of prediction; RRMSEP, relative root mean squared error of prediction; RDS, relative standard deviation.

Table 5. *Results of the simulation study when* $n = 180$ *and* $m = 40$

		$\tau^2 = 0.5$		$\tau^2 = 1$		$\tau^2 = 4$	
	σ_i^2	I	II	I	II	I	II
		Performance of the point predictor					
Bias	1	0.001	0.001	0.002	0.002	0.000	0.000
	4	0.000	0.001	0.000	0.000	0.002	0.001
	16	0.000	0.000	0.003	0.004	0.002	0.002
MSEP	1	0.352	0.374	0.514	0.531	0.808	0.813
	4	0.473	0.487	0.831	0.852	2.043	2.064
	16	0.517	0.514	0.972	0.976	3.267	3.296
		Performance of estimated MSEP					
RB	1	0.005	0.119	0.031	0.053	0.057	0.007
	4	−0.006	0.195	0.010	0.081	0.017	0.009
	16	−0.007	0.245	0.018	0.115	0.015	0.016
RRMSE	1	0.079	0.383	0.120	0.230	0.194	0.190
	4	0.036	0.578	0.058	0.332	0.122	0.158
	16	0.026	0.685	0.042	0.413	0.072	0.177
RSD	1	0.078	0.364	0.116	0.224	0.185	0.190
	4	0.035	0.544	0.057	0.322	0.121	0.158
	16	0.025	0.640	0.038	0.397	0.070	0.176

The upper part presents the bias of $\hat{\theta}_i$ and the MSEP of $\hat{\theta}_i$, whereas the lower part represents the relative bias, the relative root mean squared error and the relative standard deviation of estimated MSEP. The proposed method and the method of Wang and Fuller are abbreviated as I and II, respectively.
RB, relative bias; MSEP, mean squared error of prediction; RRMSEP, relative root mean squared error of prediction; RDS, relative standard deviation.

for $i = 1, \cdots, n$, where $E\left\{\widehat{MSEP}_i\right\}$ was estimated empirically as the average of values of $\widehat{MSEP}_i$ over 10 000 replications. The $MSEP_i$'s are as defined earlier and whose values are given in the top half part in Tables 2–5. The relative root MSE of the MSEP estimator was taken to be

$$RRMSEP_i = \frac{\left[E\left\{\widehat{MSEP}_i - MSEP_i\right\}^2\right]^{\frac{1}{2}}}{MSEP_i},$$

for $i = 1, \cdots, n$. We also report the 'relative standard deviation' (RSD) of MSEP estimators, and it is defined as $RSD = \left[Var(\widehat{MSEP}\right]^{.5} / MSEP = [RRMSEP^2 - RB^2]^{.5}$. Note that this is not exact definition of coefficient of variation, but are meant to compare the RB relative with relative root MSE. Thanks to the reviewer for suggesting this measure. The MSEP, as calculated here, is unconditional because it is not clear how to compute a conditional MSEP.

The RB of MSEP estimators under the proposed method are generally less than that of method II except the case when $\tau^2 = 4$ and $\sigma_i^2 = 1$ (high model variance, low sampling variance). Note that, in this case, the regression model is less reliable. The proposed method might have low underestimation (about 8%) in case of $n = 36$. On the other hand, method II can have very large overestimation which could be as large as 47% for small n and high ratio of sampling variance to model variance. This kind of overestimation was also noted by Rivest & Vandal (2003) when model parameters are estimated by methods of moments. Apparently, this characteristic depends on the number of small areas instead of within area sample sizes. In terms of relative root MSE of MSEP, the proposed method outperforms method II. The RSDs of MSEP estimators under the proposed method are also lower compared with that of method II, indicating more stability of our MSEP estimators. The gain is significant in most cases except the case with high model variance compared with sampling variance. Again, in this case, model is generally less reliable. The decreasing values of the RB and the relative root MSE with increasing sample size indicate consistency properties of the proposed MSEP estimators.

Rivest & Vandal (2003) extended the Prasad–Rao MSEP estimators when the direct area variances are estimated and compared their biases when the estimation of sampling variances is accounted for versus when it is not. For the sake of completeness, we report the equation here with the notations used in this paper. Equation (2.7) of Rivest & Vandal (2003) is

$$mse_{PRg}\left(\hat{\theta}_i\right) = \frac{S_i^2\hat{\tau}^2}{S_i^2 + \hat{\tau}^2} + \frac{S_i^4 Z_i^T \hat{A}^{-1} Z_i}{\left(S_i^2 + \hat{\tau}^2\right)^2} + 2\frac{S_i^4 \hat{Var}\left(\hat{\tau}^2\right) + \hat{\tau}^4\hat{\gamma}_i}{\left(S_i^2 + \hat{\tau}^2\right)^3},$$

where $\hat{\gamma}_i$ is an estimator of $\gamma_i = 2\sigma_i^4/(n_i - 1)$, $\hat{A} = \sum_i Z_i Z_i^T / \left(S_i^2 + \hat{\tau}^2\right)$ and $\hat{Var}\left(\hat{\tau}^2\right)$ is an estimator of the variance of $\hat{\tau}^2$. Note that, the Prasad–Rao MSEP estimators can be obtained by plugging in $\hat{\gamma} = 0$ in this equation.

As pointed out by a referee, it may be interesting to see the performance of Prasad–Rao MSEP estimators in the context of our simulation set-up. We computed the Prasad–Rao MSEP estimates under the following ways: (i) use true σ_i^2, but estimate β and τ^2 as described in Rivest & Vandal (2003); (ii) replace σ_i^2 by the direct estimate S_i^2, but estimate β and τ^2 as described in Rivest & Vandal (2003); and (iii) estimate σ_i^2, β and τ^2 as described in Section 3. Note that the shrinkage estimate of σ_i^2 is used in case (iii). The maximum underestimation occurs when $\left(m, n, \tau^2, \sigma_i^2\right) = (9, 36, 4, 16)$, and they are −5%, −15% and −14% for cases (i), (ii) and (iii), respectively. Thus, the Prasad–Rao MSEP estimators perform much better when the actual sampling variance is known compared with the situation when it is estimated. Note that in this scenario, our method does not have any underestimation. The RB is 3.8% . This fact supports that our method is effective when the sampling variances are unknown and only their direct estimates are available.

6. Concluding remarks

In this article, we have considered joint modelling of mean and variances for the well-known Fay–Herriot model. We derived the formula for prediction error by asymptotic expansion. The method is computationally less intensive compared with resampling-based or Markov chain Monte Carlo-based full Bayesian approach. Our approach does not need any prior specifications. The code is developed using the software R and is available upon request. In principle, the proposed techniques of prediction error calculation could be useful to obtain prediction errors for high-dimensional data analysis. For example, Hwang *et al.* (2009) used a veriant of this model for analysing microarray gene expression data.

Acknowledgments

The authors greatly acknowledge the referee comments that lead to a significant improvement of this article. This research was partially supported by NSF grants SES-0961618, SES-0961649 and DMS 1106450.

References

Arora, V. & Lahiri, P. (1997). On the superiority of the Bayesian methods over the BLUP in small area estimation problems. *Stat. Sinica* **7**, 1053–1063.

Bell, W. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 327–334.

Booth, J. & Hobert, J. (1998). Standard errors of prediction in the generalized linear mixed models. *J. Amer. Statist. Assoc.* **93**, 262–272.

Cho, M., Eltinge, J., Gershunskaya, J. & Huff, L. (2002). Evaluation of generalized variance function estimators for the U.S. current employment survey. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 534–539.

Cox, D. R. & Snell, E. J. (1968). A general definition of residuals. *J. R. Stat. Soc. Ser. B* **30**, 248–275.

Dass, S. C., Maiti, T., Ren, H. & Sinha, S. (2012). Confidence interval estimation of small area parameters shrinking both means and variances. *Surv. Methodol.* **38**, 173–187.

Eltinge, J., Cho, M. & Hinrichs, P. (2002). Use of generalized variance functions in multivariate analysis. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 904–912.

Gershunskaya, J. & Lahiri, P. (2005). Variance estimation for domains in the U.S. current employment statistics program,; 3044–3051.

Ghosh, M. & Rao, J. N. K. (1994). Small arear estimation: an appraisal. *Statist. Sci.* **9**, 54–76.

Hall, P. & Maiti, T. (2006). Nonparametric estimation of mean squared prediction error in nested-error regression moels. *Ann. Statist.* **34**, 1733–1750.

Huff, L., Eltinge, J. & Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the U.S. current employment survey. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 1519–1524.

Hwang, J. T. G., Qiu, J. & Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both mean and variances. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71**, 265–285.

Jiang, J., Lahiri, P. & Wan, S.-M. (2002). A unified Jackknife theory for empirical best prediction with M-estimation. *Ann. Stat.* **30**, 1782–1810.

Liu, B., Lahiri, P. & Kalton, G. (2007). Hierarchical Bayes modelling of survey-weighted small area proportions. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 3181–3186.

Lohr, S. & Rao, J. N. K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika* **96**, 457–468.

Maples, J., Bell, W. & Huang, E. (2009). Small area variance modeling with application to county poverty estimates from the American community survey. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 5056–5067.

Otto, M. C. & Bell, W. R. (1995). Sampling error modelling of poverty and income statistics for states. In *Proceedings of the Section on Government Statistics, American Statistical Association*, 160–165.

Pfeffermann, D. (2002). Small area estimation - new developments and directions. *Int. Statist. Rev.* **70**, 125–143.

Rao, J. N. K. (2003). Some new developments in small area estimation. *J. Iran. Stat. Soc.* **2**, 145–169.

Rivest, L.-P. & Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. In *Proceedings of the International Conference on Recent Advances in Survey Sampling*.

Wang, J. & Fuller, W. (2003). The mean squared error of small area predictors constructed with estimated error variances. *J. Amer. Statist. Assoc.* **98**, 716–723.

You, Y. & Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Surv. Methodol.* **32**, 97–103.

Received June 2011, in final form September 2013

Tapabrata Maiti, Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824.
E-mail:maiti@stt.msu.edu

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

Appendix S1 : Estimation of the parameters,
Appendix S2 : Computation of $\hat{\boldsymbol{B}} - \boldsymbol{B}$,
Appendix S3 : Second order correction of ν_i,
Appendix S4 : Information Matrix of $\boldsymbol{B}$.