Small area estimation with mixed models: a review

Shonosuke Sugasawa¹ · Tatsuya Kubokawa²

Received: 22 December 2019 / Accepted: 19 March 2020

© The Author(s) 2021, Corrected publication 2021

Abstract

Small area estimation is recognized as an important tool for producing reliable estimates under limited sample information. This paper reviews techniques of small area estimation using mixed models, covering from basic to recently proposed advanced ones. We first introduce basic mixed models for small area estimation, and provide several methods for computing mean squared errors and confidence intervals which are important for measuring uncertainty of small area estimators. Then we provide reviews of recent development and techniques in small area estimation. This paper could be useful not only for researchers who are interested in details on the methodological research in small area estimation, but also for practitioners who might be interested in the application of the basic and new methods.

Keywords Best linear unbiased predictor · Empirical Bayes · Fay–Herriot model · Hierarchical Bayes · Linear mixed model · Maximum likelihood estimator · Mean squared error · Nested error regression model · Shrinkage

1 Introduction

The terms ‘small area’ or ‘small domain’ refers to a small geographical region such as a county, municipality or state, or a small demographic group such as a specific age–sex–race group. In the estimation of a characteristic of such a small group, the direct estimate based on only the data from the small group is likely to be unreliable, because only the small number of observations are available from the small group. The problem of small area estimation is how to produce a reliable estimate for the characteristic of the small group, and the small area estimation has been actively and extensively studied from both theoretical and practical aspects due to an increasing demand for reliable small area estimates from public and private sectors. The

✉ Tatsuya Kubokawa
tatsuya@e.u-tokyo.ac.jp

¹ Center for Spatial Information Science, The University of Tokyo, Tokyo, Japan

² Faculty of Economics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

articles Ghosh and Rao (1994) and Pfeffermann (2013) give good reviews and motivations, and the comprehensive book Rao and Molina (2015) covers all the main developments in small area estimation. Also see Demidenko (2004) for general mixed models and Pratesi (2016) for the analysis of poverty data by small area estimation. In this paper, we describe the details of classical methods and give a review of recent developments, which will be helpful for readers who are interested in this topic.

To improve the accuracy of direct survey estimates, we make use of the relevant supplementary information such as data from other related areas and covariate data from other sources. The linear mixed models (LMM) enable us to ‘borrow strength’ from the relevant supplementary data, and the resulting model-based estimators or the best linear unbiased predictors (BLUP) provide reliable estimates for the small area characteristics. The BLUP shrinks the direct estimates in small areas towards a stable quantity constructed by pooling all the data, thereby BLUP is characterized by the effects of pooling and shrinkage of the data. These two features of BLUP mainly come from the structure of linear mixed models described as (observation) = (common parameters) + (random effects) + (error terms), namely the shrinkage effect arises from the random effects, and the pooling effect is due to the setup of the common parameters. While BLUP was originally proposed by Henderson (1950), empirical version of BLUP (EBLUP) is related to the classical shrinkage estimator studied by Stein (1956), who established analytically that EBLUP improves on the sample means when the number of small areas is larger than or equal to three. This fact shows not only that EBLUP has a larger precision than the sample mean, but also that a similar concept came out at the same time by Henderson (1950) for practical use and Stein (1956) for theoretical interest. Based on these historical backgrounds, there have been a lot of methods proposed so far. In Sect. 2, we describe the details of classical approaches to small area estimation through the two basic linear mixed models: the Fay–Herriot (FH) model suggested by Fay and Herriot (1979) and the nested error regression (NER) model, which are used for the analysis of area-level and unit-level data, respectively. These models are classical, but helpful for understanding how to make reliable estimates. The Fay–Herriot model is extended to the unmatched sampling and linking models for analyzing positive data and hierarchical models based on natural exponential families for analyzing area-level discrete data such as mortality. These extended models are also explained in Sect. 2.

In Sect. 3, we discuss the uncertainty quantification of small area estimators. Since the small area estimator is constructed through model-based approaches, it is quite important in practice to assess the variability or risk for the resulting estimates. We mainly focus on mean squared errors and prediction intervals, and describe several methods based on asymptotic calculations and simulation-based methods such as jackknife and bootstrap.

The classical and basic results in small area estimation are explained in Sects. 2 and 3. Due to a growing demand for reliable small area statistics, however, the small area estimation has been extensively studied in various directions from both theoretical and practical aspects. In Sect. 4, we give a review of recent developments. The topics treated there are adjusted likelihood methods for variance components estimation, estimation of general parameters such as poverty measures in a finite

population, transforming response values via parametric transformation, flexible modeling for heteroscedastic variances, modeling of random effects based on the spike-and-slab and global–local shrinkage priors, skewed distributions for error terms and random effects, nonparametric and semiparametric modeling, robust methods, measurement error models, observed best prediction methods, and variable selection techniques. These explanations will be helpful for readers interested in small area estimation. This section shows that practical situations in real data analysis request new methodology, techniques and theories, and the small area estimation is developing from practical and theoretical points of view. The paper is concluded with some remarks in Sect. 5. Since mainly recent articles are given in the reference of this paper, see the reference in Rao and Molina (2015) for other articles.

2 Basic mixed models for small area estimation

In this section, we introduce some basic mixed models used in small area estimation. We start with two most standard models, known as the Fay–Herriot model and the nested error regression model. We also describe some models for non-normal data and sub-area models.

2.1 Fay–Herriot model

Most public data are reported based on cumulated data like sample means from counties and cities. The Fay–Herriot (FH) model introduced by Fay and Herriot (1979) is the mixed model for estimating the true areal means $\theta_1, \dots, \theta_m$ based on area-level summary statistics denoted by $y_1, \dots, y_m$, where y_i is called a direct estimator of θ_i for $i = 1, \dots, m$. Note that y_i is a crude estimator θ_i with large variance, because the sample size for calculating y_i could be small in practice. Let $\mathbf{x}_i$ be a vector of known area characteristics with an intercept term. The FH model is given by

$$y_i = \theta_i + \varepsilon_i, \quad \theta_i = \mathbf{x}_i' \boldsymbol{\beta} + v_i, \quad i = 1, \dots, m, \quad (2.1)$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients, ε_i and v_i are sampling errors and random effects, respectively, and are independently distributed as $\varepsilon_i \sim N(0, D_i)$ and $v_i \sim N(0, A)$. Here D_i is a variance of y_i given θ_i , which is assumed to be known, and A is an unknown variance parameter. The assumption that D_i is known seems restrictive, because it can be estimated from data a priori. This issue will be addressed in Sect. 4.4.

The best predictor of θ_i under squared error loss is the conditional expectation given by

$$E[\theta_i | y_i] = \gamma_i y_i + (1 - \gamma_i) \mathbf{x}_i' \boldsymbol{\beta}, \quad (2.2)$$

where $\gamma_i = A/(A + D_i)$ is known as a shrinkage coefficient. Under model (2.1), the marginal distribution of y_i is $N(\mathbf{x}_i' \boldsymbol{\beta}, A + D_i)$, and $\boldsymbol{\beta}$ can be estimated under given A by generalized least squares (GLS) estimator given by

$$\hat{\beta}_{\text{GLS}} = \left(\sum_{i=1}^m \frac{\mathbf{x}_i \mathbf{x}_i'}{A + D_i} \right)^{-1} \left(\sum_{i=1}^m \frac{\mathbf{x}_i y_i}{A + D_i} \right). \quad (2.3)$$

By replacing β with $\hat{\beta}_{\text{GLS}}$ in (2.2), we obtain the following best linear unbiased predictor (BLUP):

$$\tilde{\theta}_i \equiv \gamma_i y_i + (1 - \gamma_i) \mathbf{x}_i' \hat{\beta}_{\text{GLS}}. \quad (2.4)$$

Since $\hat{\beta}_{\text{GLS}}$ is constructed based on all the data, the regression estimator $\mathbf{x}_i' \hat{\beta}_{\text{GLS}}$ would be much more stable than the direct estimators y_i . Then the BLUP (2.4) can be interpreted as a shrinkage estimator that shrinks the unstable direct estimator y_i toward the stable estimator $\mathbf{x}_i' \hat{\beta}_{\text{GLS}}$, depending on the shrinkage coefficient γ_i . Note that if D_i is large compared with A , which means that y_i has a large fluctuation, γ_i is small, so that y_i is more shrunk toward $\mathbf{x}_i' \hat{\beta}_{\text{GLS}}$, and vice versa.

In practice, the random effects variance A is unknown, and should be replaced in γ_i and $\hat{\beta}_{\text{GLS}}$ by a sample estimate, which yields the empirical BLUP (EBLUP) in the frequentist's framework, or the empirical Bayes estimator in the Bayesian framework. The standard way to estimate A is the maximum likelihood estimator based on the marginal distribution of y_i . Other options would be the restricted maximum likelihood estimator and moment-type estimators as considered in Fay and Herriot (1979) and Prasad and Rao (1990). We will revisit this issue in Sect. 4.1. Alternatively, we may employ the hierarchical Bayes (HB) approach by assigning prior distributions on unknown parameters β and A , and compute a posterior distribution of θ_i , which produces the point estimator as well as credible intervals. Due to the recent advancement of computational techniques, the HB approaches now became standard in this context (Rao and Molina 2015).

2.2 Nested error regression model

When unit-level data are available, we can use a model for more in-depth analysis. Let $y_{i1}, \dots, y_{in_i}$ be a unit-level sample from the i th area for $i = 1, \dots, m$, and let $\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}$ be fixed vectors of covariates with/without the intercept term. The nested error regression model (Battese et al. 1988) is described as

$$y_{ij} = \mathbf{x}_{ij}' \beta + v_i + \varepsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, m, \quad (2.5)$$

where v_i and ε_{ij} are random effects and error terms, respectively, and are mutually independently distributed as $v_i \sim N(0, \tau^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$, β is a vector of unknown regression coefficients, and τ^2 and σ^2 are unknown variance parameters. It is noted that v_i is a random effect depending on the i th area and common to all the observations y_{ij} 's in the same area. This induces correlations among y_{ij} 's, that is, $\text{Cov}(y_{ij}, y_{ij'}) = \tau^2$ for $j \neq j'$, noting that observations in the different areas are still independent. Hence, the variances σ^2 and τ^2 are called 'within' and 'between' components of variance, respectively.

The NER model is typically used in the framework of a finite population model. We assume that the i th area includes N_i units in total, but only n_i units are sampled. For simplicity, we assume sampling mechanism is the simple random sampling, so that we do not consider survey weights. For all the units, we assume the following super-population model:

$$Y_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad j = 1, \dots, N_i, \quad i = 1, \dots, m,$$

where Y_{ij} is the characteristic for the j th unit in the i th area. Without loss of generality, we assume the first n_i characteristics $Y_{i1}, \dots, Y_{in_i}$ are observed as $y_{i1}, \dots, y_{in_i}$, and the rest of characteristics $Y_{i,n_i+1}, \dots, Y_{iN_i}$ are unobserved. Under the setting, the true area mean is defined as

$$\frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij} = \frac{1}{N_i} \sum_{j=1}^{N_i} (\mathbf{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij}) = \bar{\mathbf{X}}_i^t \boldsymbol{\beta} + v_i + \frac{1}{N_i} \sum_{j=1}^{N_i} \varepsilon_{ij}.$$

In practice, the total number of units N_i is very large although the number of sampled units n_i is not large. Then the final term could be negligible, thereby we can define the mean parameter θ_i as $\theta_i = \bar{\mathbf{X}}_i^t \boldsymbol{\beta} + v_i$. Hence, we can estimate θ_i only if we know the true mean vector $\bar{\mathbf{X}}_i$ of auxiliary information, which is actually often the case in practice. Under the model (2.5), the best predictor of v_i is given by

$$\tilde{v}_i = \frac{n_i \tau^2}{\sigma^2 + n_i \tau^2} (\bar{y}_i - \bar{\mathbf{x}}_i^t \boldsymbol{\beta}), \quad (2.6)$$

where $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ and $\bar{\mathbf{x}}_i = n_i^{-1} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$. Similar to the FH model, $\boldsymbol{\beta}$ can be estimated via the GLS estimator $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ based on all the sampled data. Also variance parameters τ^2 and σ^2 can be estimated via the marginal distribution of $(y_{i1}, \dots, y_{in_i})$, which is Gaussian under (2.5). Then the EBLUP $\hat{v}_i$ of v_i can be obtained, so that the EBLUP of θ_i is given by $\hat{\theta}_i = \bar{\mathbf{x}}_i^t \hat{\boldsymbol{\beta}}_{\text{GLS}} + \hat{v}_i$. More general methods for predicting population means are discussed in Jiang and Lahiri (2006).

As a famous application of NER, Battese et al. (1988) considered predicting areas under corn and soybeans for each of $m = 12$ counties in north-central Iowa. In their analysis, each county is divided into about 250 hectares segments, and n_i segments are selected from the i th county. For the j th segment of the i th county, y_{ij} is the number of hectares of corn (or soybeans) in the (i, j) segment reported by interviewing farm operators, and x_{ij1} and x_{ij2} are the numbers of pixels (0.45 hectare) classified as corn and soybeans, respectively, using LANDSAT satellite data. Since n_i 's range from 1 to 5 with $\sum_{i=1}^m n_i = 37$, the sample mean $\bar{y}_i$ has large deviation for predicting the mean crop hectare per segment θ_i . The NERM enables us to construct more reliable prediction procedures not only using the auxiliary information on the LANDSAT data, but also by combining the data of the related areas.

2.3 Unmatched sampling and linking models

Response variables y_i for which the normality assumption is not suitable (e.g. income) are ubiquitous in real applications. In this case, we may use the generalized version of the FH model (2.1), given by

$$y_i = \theta_i + \varepsilon_i, \quad h(\theta_i) = \mathbf{x}_i^t \boldsymbol{\beta} + v_i, \quad i = 1, \dots, m, \quad (2.7)$$

where $h(\cdot)$ is a known link function, and the other variables are defined in the same as in (2.1). A typical example for positive valued y_i is $h(x) = \log x$. Under the model (2.7), the marginal distribution of y_i is expressed as

$$\int \phi(y_i; \theta_i, D_i) \phi(h(\theta_i); \mathbf{x}_i^t \boldsymbol{\beta}, A) h'(\theta_i) d\theta_i,$$

where h' is the first-order derivative of h , and $\phi(\cdot; a, b)$ denotes the density function of the normal distribution with mean a and variance b . The above density does not admit a closed form expression for a general function $h(\cdot)$. The model (2.7) is originally proposed by You and Rao (2002), who considered a hierarchical Bayesian approach via the Markov Chain Monte Carlo algorithm. Recently, Sugawara et al. (2018) developed an empirical Bayes approach in which the Monte Carlo method for computing the best predictor of θ_i and the Monte Carlo expectation-maximization algorithm for computing the maximum likelihood estimator.

2.4 Generalized linear mixed models

When the response variables are discrete such as binary, count and multi-category, discrete distributional models are required. Ghosh et al. (1998) discussed the use of generalized linear mixed models for small area estimation under this situation, and provided hierarchical Bayesian methods. On the other hand, Jiang and Lahiri (2001) discussed a frequentist approach for small area estimation with binary data. The unit-level model is given by

$$P(y_{ij} = 1 | p_{ij}) = p_{ij}, \quad \log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = \mathbf{x}_{ij}^t \boldsymbol{\beta} + u_i,$$

where $u_i \sim N(0, \tau^2)$ is a random effect. A typical purpose is to predict the true area proportion given by $\bar{p}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$. The best predictor of the random effect u_i under squared loss is the conditional expectation of u_i given $y_i = (y_{i1}, \dots, y_{in_i})$, which is expressed as

$$E[u_i | y_i] = \tau \frac{E_z[z \phi_i(z; \boldsymbol{\beta}, \tau^2)]}{E_z[\phi_i(z; \boldsymbol{\beta}, \tau^2)]},$$

where $\log \phi_i(z; \boldsymbol{\beta}, \tau^2) = \tau z \sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i} \log \{1 + \exp(\mathbf{x}_{ij}^t \boldsymbol{\beta} + \tau z)\}$, $z \sim N(0, 1)$, and E_z denotes the expectation with respect to z . The above integral does not admit

analytical expressions, thereby the numerical integration is required to compute the predictor. In general, under generalized linear mixed models, the best predictor of the random effect as well as the marginal distribution of y_i do not have analytical expressions, which could be a potential drawback of generalized linear mixed models.

2.5 Models based on natural exponential families

One of the main drawbacks of generalized linear mixed models is the intractability of conditional distributions of random effects as well as marginal likelihood functions. As alternative models for area-level discrete data, Ghosh and Maiti (2004) introduced models based on the natural exponential families. Let $y_1, \dots, y_m$ be mutually independent random variables where the conditional distribution of y_i given θ_i and the marginal distribution of θ_i belong to the following natural exponential families:

$$\begin{aligned} f(y_i|\theta_i) &= \exp[n_i\{\theta_i y_i - \psi(\theta_i)\} + c(y_i, n_i)], \\ \pi(\theta_i) &= \exp[v\{m_i \theta_i - \psi(\theta_i)\}]C(v, m_i), \end{aligned} \quad (2.8)$$

where n_i is a known scalar and v is an unknown scalar. Here $c(\cdot, \cdot)$ and $C(\cdot, \cdot)$ are normalizing constants and $\psi(\cdot)$ is a link function. Moreover, $m_i = \psi'(\mathbf{x}_i^t \boldsymbol{\beta})$, where $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are vectors of covariates and unknown regression coefficients, respectively, and $\psi'(\cdot)$ is the first order derivative of $\psi(\cdot)$. Usually, n_i is related to sample size in the i th area, so that n_i is not so large in practice. The function $f(y_i|\theta_i)$ is the regular one-parameter exponential family and the function $\pi(\theta_i)$ is the conjugate prior distribution. Note that $\mu_i \equiv E[y_i|\theta_i] = \psi'(\theta_i)$ which is the true area mean, and $E[\theta_i] = m_i$. Therefore, y_i is regarded as a crude estimator of μ_i and m_i is the prior mean of θ_i .

Owing to the conjugacy, the marginal likelihood can be obtained in a closed form, which enables us to get the maximum likelihood estimators or moment-type estimators as considered in Ghosh and Maiti (2004) for the unknown parameters. Moreover, the conditional distribution of θ_i given y_i can be obtained in an analytical form as well, and the conditional expectation of μ_i with estimated parameters is given by

$$\hat{\mu}_i = \frac{n_i y_i + \hat{v} \hat{m}_i}{n_i + \hat{v}},$$

where $\hat{m}_i = \psi'(\mathbf{x}_i^t \hat{\boldsymbol{\beta}})$. It is observed that $\hat{\mu}_i$ is a weighted average of the direct estimator y_i and the estimator of prior mean (regression estimator) $\hat{m}_i$, which would be stable since $\hat{\boldsymbol{\beta}}$ is estimated based on all the data. The model (2.8) includes some typical two-stage models for area-level data. For example, when $v = A^{-1}$ and $n_i = D_i^{-1}$ and $\psi(x) = x^2/2$, the model (2.8) reduces to the FH model (2.1). Also, when $\psi(x) = \exp(x)$ and $\psi(x) = \log(1 + \exp(x))$, the model (2.8) reduces, respectively, to the Poisson-gamma model (e.g. Clayton and Kaldor 1987) and the binomial-beta model (e.g. Williams 1975), which have been widely adopted in several fields such as disease mapping.

2.6 Some remarks

There are some papers concerned with practical applications of the standard models described above. For example, Schmid et al. (2017) employed the FH model to estimate sociodemographic indicators and Mauro et al. (2017) considered small area estimation for forest inventory. We also note that software for applying the standard mixed models for small area estimation is available recently, e.g. package ‘sae’ in +R+ language (Molina and Marhuenda 2015).

3 Measuring uncertainty of small area estimators

An important aspect of small area estimation is the assessment of the accuracy of the predictors. Under the frequentist approach, this will be complicated due to the additional fluctuation induced by estimating unknown parameters in models. We here focus on two methods that are widely adopted in this context: estimators of mean squared error (MSE) and confidence intervals.

3.1 Estimation of MSE

We first consider a general situation. Let θ_i ($i = 1, \dots, m$) be the true parameter, $\tilde{\theta}_i$ be the conditional expectation (or best predictor) of θ_i given y_i which depends on the unknown parameter $\boldsymbol{\psi}$, and $\hat{\theta}_i$ is the empirical best predictor of θ_i . MSE of $\hat{\theta}_i$ is defined by

$$\text{MSE}_i(\boldsymbol{\psi}) \equiv E[(\hat{\theta}_i - \theta_i)^2].$$

Using the fact that $\tilde{\theta}_i$ is the conditional expectation of θ_i given y_i 's, we have

$$\text{MSE}_i(\boldsymbol{\psi}) = E[(\tilde{\theta}_i - \theta_i)^2] + E[(\hat{\theta}_i - \tilde{\theta}_i)^2] \equiv g_{i1}(\boldsymbol{\psi}) + g_{i2}(\boldsymbol{\psi}). \quad (3.1)$$

Note that $g_{i1}(\boldsymbol{\psi})$ is concerned with variability of the best predictor given $\boldsymbol{\psi}$, and is typically of order $O(1)$. Also an analytical expression of $g_{i1}(\boldsymbol{\psi})$ can be obtained in many cases. On the other hand, $g_{i2}(\boldsymbol{\psi})$ measures additional variability which comes from the estimation of $\boldsymbol{\psi}$, thereby $g_{i2}(\boldsymbol{\psi}) = O(m^{-1})$ in most cases, but it hardly admits closed form expressions. Hence, we derive an approximation formula for $g_{i2}(\boldsymbol{\psi})$ up to second order.

Here we demonstrate the approximation under the FH model as considered in Datta et al. (2005). Note that $\boldsymbol{\psi} = (\boldsymbol{\beta}^t, A)^t$ in this case. From (2.2), it follows that $g_{i1}(\boldsymbol{\psi}) = AD_i/(A + D_i)$. Regarding $g_{i2}(\boldsymbol{\psi})$, it follows that

$$\hat{\theta}_i - \tilde{\theta}_i = (1 - \hat{\gamma}_i)\mathbf{x}_i^t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + (\hat{\gamma}_i - \gamma_i)(y_i - \mathbf{x}_i^t\boldsymbol{\beta}),$$

thereby it holds from Prasad and Rao (1990) and Datta and Lahiri (2000) that

$$g_{i2}(\boldsymbol{\psi}) = \frac{D_i^2}{(A + D_i)^2} \mathbf{x}_i^t \left(\sum_{k=1}^m \frac{\mathbf{x}_k \mathbf{x}_k^t}{A + D_k} \right)^{-1} \mathbf{x}_i + \frac{D_i^2}{(A + D_i)^3} \text{Var}(\hat{A}) + o(m^{-1}),$$

where $\text{Var}(\hat{A})$ is an asymptotic variance of $\hat{A}$. For more general linear mixed models, Datta and Lahiri (2000) provided the second order approximation of MSE under linear mixed models that include the FH and NER models as special cases, and Das et al. (2004) extended the results of Datta and Lahiri (2000) by relaxing the independence assumption for random effects and error terms. Moreover, Lahiri and Rao (1995) derived an approximation of MSE under the FH model without assuming normality for random effects and error terms. For non-normal models, Ghosh and Maiti (2004) derived an approximation formula of MSE under the natural exponential family model (2.8).

We now consider the estimation of MSE. In the context of small area estimation, second order unbiased estimators of MSE having bias of order $o(m^{-1})$ are widely adopted. To achieve the goal, there are mainly three methods: analytical methods (e.g. Prasad and Rao 1990; Das et al. 2004; Datta et al. 2005; Lahiri and Rao 1995), bootstrap methods (e.g. Butar and Lahiri 2003; Hall and Maiti 2006a, b), and Jackknife methods (e.g. Jiang et al. 2002), which are described as follows.

(1) Analytical method We assume $g_{i1} = O(1)$ and $g_{i2} = O(m^{-1})$ in (3.1). Then the plug-in estimator $g_{i2}(\hat{\boldsymbol{\psi}})$ is second-order unbiased, whereas $g_{i1}(\hat{\boldsymbol{\psi}})$ has a second-order bias. The expectation $E[g_{i1}(\hat{\boldsymbol{\psi}})]$ is approximated as

$$E[g_{i1}(\hat{\boldsymbol{\psi}})] = g_{i1}(\boldsymbol{\psi}) + \frac{\partial g_{i1}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^t} E[\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}] + \frac{1}{2} \text{tr} \left\{ \text{Var}(\hat{\boldsymbol{\psi}}) \left(\frac{\partial^2 g_{i1}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^t} \right) \right\} + o(m^{-1}).$$

For simplicity, let $B_i(\boldsymbol{\psi}) = E[g_{i1}(\hat{\boldsymbol{\psi}})] - g_{i1}(\boldsymbol{\psi})$. Then we can define a bias corrected estimator of g_{i1} , that is, $g_{i1}(\hat{\boldsymbol{\psi}}) - B_i(\hat{\boldsymbol{\psi}})$, which is second-order unbiased under some regularity conditions. Therefore, an analytical estimator of MSE is given by

$$\widehat{\text{MSE}}_i = g_{i1}(\hat{\boldsymbol{\psi}}) - B_i(\hat{\boldsymbol{\psi}}) + g_{i2}(\hat{\boldsymbol{\psi}}), \quad (3.2)$$

which satisfies $E[\widehat{\text{MSE}}_i] = \text{MSE}_i + o(m^{-1})$. To derive the MSE estimator, we derive an asymptotic bias of $\hat{\boldsymbol{\psi}}$, which typically requires tedious algebraic calculation. It is noted that the MSE estimator (3.2) does not necessarily produce positive-valued estimators because of the bias correction although the true MSE values are positive.

(2) Bootstrap method There are a few types of bootstrap methods for MSE. The most typical approach would be a hybrid bootstrap method (Butar and Lahiri 2003) which separately estimates g_{i1} and g_{i2} via the parametric bootstrap. Here we explicitly write $\tilde{\theta}_i(y_i, \boldsymbol{\psi})$ instead of $\tilde{\theta}_i(\boldsymbol{\psi})$ in order to address the dependence of the best predictor $\tilde{\theta}_i$ on the observation y_i . Given the estimate $\hat{\boldsymbol{\psi}} = \hat{\boldsymbol{\psi}}(y_1, \dots, y_m)$ of $\boldsymbol{\psi}$, the parametric bootstrap method first generates the bootstrap sample $y_{1(b)}^*, \dots, y_{m(b)}^*$ from the assumed model with $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}$ as the b th bootstrap sample, and then computes the bootstrap estimator $\hat{\boldsymbol{\psi}}_b^* = \hat{\boldsymbol{\psi}}^*(y_{1(b)}^*, \dots, y_{m(b)}^*)$ of $\boldsymbol{\psi}$. Then the hybrid bootstrap estimator is given by

$$2g_{i1}(\hat{\psi}) - \frac{1}{B} \sum_{b=1}^B g_{i1}(\hat{\psi}_b^*) + \frac{1}{B} \sum_{b=1}^B \left\{ \tilde{\theta}_i(y_{i(b)}^*, \hat{\psi}_b^*) - \tilde{\theta}_i(y_i, \hat{\psi}_b^*) \right\}^2,$$

where B is the number of bootstrap replications. Another bootstrap method is the double bootstrap (e.g. Hall and Maiti 2006b) in which the naive bootstrap estimator is defined as

$$M(y^*) = \frac{1}{B} \sum_{b=1}^B \left\{ \tilde{\theta}_i(y_{i(b)}^*, \hat{\psi}_b^*) - \theta_{i(b)}^* \right\}^2,$$

where $\theta_{i(b)}^*$ is a generated value of θ_i from the estimated model and y^* is a collection of the bootstrap sample. Then the double bootstrap estimator computes C bootstrap MSE estimators $\{M(y_1^*), \dots, M(y_C^*)\}$ and carry out bias correction. Compared with the hybrid bootstrap estimator, the double bootstrap method requires additional bootstrap replications, which could be computationally intensive in practice.

(3) Jackknife method Jiang et al. (2002) suggested the use of the jackknife method for estimating g_{i1} and g_{i2} , separately. Let $\hat{\psi}_{-\ell}$ denote the estimator of ψ based on all the data except for the ℓ th area. Then the Jackknife estimator of MSE is given by $\hat{g}_{i1}^J + \hat{g}_{i2}^J$, where

$$\begin{aligned} \hat{g}_{i1}^J &= g_{i1}(\hat{\psi}) - \frac{m-1}{m} \sum_{\ell=1}^m \{g_{i1}(\hat{\psi}_{-\ell}) - g_{i1}(\hat{\psi})\}, \\ \hat{g}_{i2}^J &= \frac{m-1}{m} \sum_{\ell=1}^m \{\tilde{\theta}_i(\hat{\psi}_{-\ell}) - \tilde{\theta}_i(\hat{\psi})\}^2. \end{aligned}$$

Under some regularity conditions, it holds that the estimator is second-order unbiased.

Finally, we note that there is another type of MSE, called conditional MSE, defined as $E[(\hat{\theta}_i - \theta_i)^2 | y_i]$, which measures the estimation variability under given y_i . The detailed investigation and comparisons with the standard (unconditional) MSE have been done in the literature (e.g. Datta et al. 2011; Torabi and Rao 2013). As noted in Booth and Hobert (1998) and Sugawara and Kubokawa (2016), the difference between the conditional and unconditional MSEs under models based on normal distributions can be negligible under large sample sizes (i.e. large number of areas), whereas the difference is significant under non-normal distributions. Also unified jackknife methods for the conditional MSE are developed in Lohr and Rao (2009).

3.2 Confidence intervals

Another approach to measuring uncertainty of EBLUP is a confidence interval based on EBLUP, and the confidence intervals which satisfy the nominal confidence level with second-order accuracy are desirable. There are mainly two methods for

constructing the confidence interval: the analytical method based on a Taylor series expansion and a parametric bootstrap method.

For simplicity, we suppose that $\theta_i|y_i \sim N(\tilde{\theta}_i(y_i, \boldsymbol{\psi}), s_i(\boldsymbol{\psi})^2)$, where $\tilde{\theta}_i$ and s_i^2 are conditional expectation and variance of θ_i . This holds under the FH model (2.1) or general class of linear mixed models. Then, given the unknown parameter $\boldsymbol{\psi}$, the confidence interval with coverage probability $1 - \alpha$ is $I_\alpha(\boldsymbol{\psi}) : \tilde{\theta}_i(y_i, \boldsymbol{\psi}) \pm z_{\alpha/2}s_i(\boldsymbol{\psi})$ with z_α being the upper $100\alpha\%$ quantile of the standard normal distribution, which satisfies $P(\theta_i \in I_\alpha(\boldsymbol{\psi})) = 1 - \alpha$. Since $\boldsymbol{\psi}$ is unknown, the naive confidence interval is $I_\alpha(\hat{\boldsymbol{\psi}})$, which has a coverage probability with the approximation $P(\theta_i \in I_\alpha(\hat{\boldsymbol{\psi}})) = 1 - \alpha + O(m^{-1})$ in general. Thus, a confidence interval with the second-order correction $P(\theta_i \in I_\alpha(\hat{\boldsymbol{\psi}})) = 1 - \alpha + O(m^{-3/2})$ is desirable.

(1) Analytical method We begin by expanding the coverage probability of the naive interval. For example, under the FH model (2.1), we have

$$P(\theta_i \in I_\alpha(\hat{\boldsymbol{\psi}})) = 2\Phi(z_{\alpha/2}) - 1 - z_{\alpha/2}\phi(z_{\alpha/2})c_i(z_{\alpha/2}, A) + O(m^{-3/2}),$$

where

$$c_i(z, A) = \left\{ \sum_{i=1}^m \frac{1}{(A + D_i)^2} \right\}^{-1} \left\{ \frac{4D_i}{A(A + D_i)^2} + \frac{(1 + z^2)D_i^2}{2A^2(A + D_i)^2} \right\} + \frac{D_i}{A(A + D_i)} \mathbf{x}_i' \text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) \mathbf{x}_i,$$

noting that $c_i(z, A) = O(m^{-1})$ (e.g. Yoshimori and Lahiri 2014b). Then the confidence interval obtained by replacing $z_{\alpha/2}$ with $z_{\alpha/2}\{1 + c_i(z, A)\}$ has the corrected coverage probability $1 - \alpha + O(m^{-3/2})$. Similar results are obtained in Datta et al. (2002) and Basu et al. (2003).

As another approach, we may construct confidence interval-based MSE estimators given in the previous section, which has the form

$$\tilde{\theta}(y_i, \hat{\boldsymbol{\psi}}) \pm z_{\alpha/2}\{1 + h_i(\hat{\boldsymbol{\psi}})\}\widehat{\text{MSE}}_i^{1/2},$$

where $\widehat{\text{MSE}}_i$ is a second-order unbiased MSE estimator, and $h_i(\cdot)$ is an adjustment function such that the coverage of the above interval is second-order accurate. Diao et al. (2014) adopted this idea to derive confidence intervals under the FH model. Also Kubokawa (2009) derived the confidence intervals of this type under the NER model.

(2) Bootstrap method We first provide a method using a pivotal statistic given in Chatterjee et al. (2008). Define $U_i(\boldsymbol{\psi}) = (\theta_i - \tilde{\theta}_i)/s_i(\boldsymbol{\psi})$, then $U_i(\boldsymbol{\psi}) \sim N(0, 1)$ when $\boldsymbol{\psi}$ is the true parameter. We approximate the distribution of $U_i(\hat{\boldsymbol{\psi}})$ via the parametric bootstrap, that is, we generate the parametric bootstrap sample $\theta_{i(b)}^*$ as well as $y_{i(b)}^*$ from the estimated model and compute the bootstrap estimator $\hat{\boldsymbol{\psi}}_{(b)}^*$ for $b = 1, \dots, B$. Then the distribution of $U_i(\hat{\boldsymbol{\psi}})$ can be approximated by B bootstrap realizations $\{U_{i(b)}^*, b = 1, \dots, B\}$, where $U_{i(b)}^* = (\theta_{i(b)}^* - \hat{\theta}_{i(b)}^*)/s_i(\hat{\boldsymbol{\psi}}_{(b)}^*)$. Letting $z_{iu}^*(\alpha)$ and $z_{il}^*(\alpha)$ be the empirical upper and lower $100\alpha\%$ quantiles of the empirical distribution of

$\{U_{i(b)}^*, b = 1, \dots, B\}$, the calibrated confidence interval is given by $(\hat{\theta}_i + z_{il}^*(\alpha/2)s_i(\hat{\psi}), \hat{\theta}_i + z_{iu}^*(\alpha/2)s_i(\hat{\psi}))$.

We next describe a general parametric bootstrap approach given in Hall and Maiti (2006b). Define $I_\alpha(\psi) = (F_{\alpha/2}(\psi), F_{1-\alpha/2}(\psi))$ where $F_\alpha(\psi)$ is the α -quantile of the posterior distribution of θ_i , such as $N(\hat{\theta}_i(y_i, \psi), s_i(\psi)^2)$. Since the naive interval $I_\alpha(\hat{\psi})$ does not satisfy $P(\theta_i \in I_\alpha(\hat{\psi})) = 1 - \alpha$, we calibrate a suitable α via the parametric bootstrap. Denote by $\hat{I}_{\alpha(b)}^* = I_\alpha(\hat{\psi}_{(b)}^*)$ the bootstrap interval based on the b th bootstrap sample, and let $\hat{\alpha}$ be the solution of the equation $B^{-1} \sum_{b=1}^B I(\theta_i^* \in \hat{I}_{\alpha(b)}^*) = \alpha$. Then, $I_{\hat{\alpha}}(\hat{\psi})$ is the bootstrap-calibrated interval which has a coverage probability with second-order accuracy.

4 Recent developments in small area estimation

The small area estimation has been actively and extensively studied in various directions from theoretical and practical aspects. In this section, we give a review of recent developments, which will be helpful for readers interested in this topic.

4.1 Adjusted likelihood method

Remember that the random effects variance A determines the amount of shrinkage through γ_i in the BLUP (2.4). A practical problem in estimating A is that the maximum likelihood estimator may produce zero estimate, under which the BLUP (2.4) reduces to $\tilde{\theta}_i = \mathbf{x}_i' \hat{\beta}_{\text{GLS}}$. This is not plausible since heterogeneity among areas cannot be taken into account. To overcome the difficulty, Li and Lahiri (2010) introduced a modification of the likelihood function called an adjusted likelihood, given by

$$L_{\text{ad}}(A) = h(A)L(A), \quad (4.1)$$

where $L(A)$ is the profile likelihood of A and $h(A)$ is an adjustment factor. The function $h(A)$ should be chosen such that the adjusted likelihood $L_{\text{ad}}(A)$ tends to be 0 around $A = 0$. Specifically, Li and Lahiri (2010) employed $h(A) = A$, and derived asymptotic properties of the resulting estimator. Moreover, Yoshimori and Lahiri (2014a) proposed an alternative form, $h(A) = [\tan^{-1}\{\text{tr}(\mathbf{I}_m - \mathbf{B})\}]^{1/m}$ with $\mathbf{B} = \text{diag}(D_1/(A + D_1), \dots, D_m/(A + D_m))$, to make the resulting estimator $\hat{A}$ hold some desirable asymptotic properties.

Such an adjusted likelihood method has been adopted not only for avoiding zero estimate but also for MSE estimation and constructing confidence intervals since the adjustment likelihood method produces different asymptotic properties of $\hat{A}$. Yoshimori and Lahiri (2014b) is concerned with constructing efficient confidence intervals with second-order accuracy via the adjusted likelihood method. Under the FH model (2.1), the authors focused on the confidence interval of θ_i of the form, $I_i(A) : \tilde{\theta}_i \pm z_{\alpha/2} \sqrt{AD_i/(A + D_i)}$, where $\tilde{\theta}_i$ is given in (2.2) and z_α is the upper $100\alpha\%$ quantile of the standard normal distribution. Yoshimori and Lahiri (2014b) considered an area-specific adjustment function $h_i(A)$ to produce the area-specific estimator $\hat{A}_i$. Then they derived a condition on $h_i(A)$ such that the coverage probability of

the confidence interval $I_i(\hat{A}_i)$ has second-order accuracy. Moreover, Hirose (2017) improved the result so that the adjustment function does not depend on i . Regarding the MSE estimation, Hirose (2019) employed the same idea to derive a second-order unbiased estimator of MSE which is always positive under the FH model (2.1). The author considered a certain class for the adjustment factor and gave reasonable conditions to achieve the desirable properties of MSE estimation. Finally, Hirose and Lahiri (2018) proposed a new but simple area-wise adjustment factor $h_i(A) = A + D_i$ to achieve several nice properties simultaneously.

4.2 Estimation of general parameters in a finite population

In Sect. 2.2, we demonstrated an application of the NER model to estimating areal means in the finite population framework. Molina and Rao (2010) extended the classical approach to the estimation of the general parameters given by

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} T(Y_{ij}),$$

where $T(\cdot)$ depends on estimation problems which we are interested in. For example, if we adopt $T(x) = I(x < z)$ for a fixed thresholding value z and Y_{ij} is a welfare measure, then μ_i can be interpreted as the poverty rate in the i th area. Molina and Rao (2010) introduced the FGT poverty measure (Foster et al. 1984) to estimate area-wise general poverty indicators. Molina and Rao (2010) used the following NER model:

$$H(Y_{ij}) = \mathbf{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad j = 1, \dots, N_i, \quad i = 1, \dots, m, \quad (4.2)$$

where v_i and ε_{ij} are defined in the same way as in Sect. 2.2, and $H(\cdot)$ is a specified transformation function such as the logarithm function. Here we again assume that the first n_i units are sampled without loss of generality, and let $s_i = \{1, \dots, n_i\}$ and $r_i = \{n_i + 1, \dots, N_i\}$. Under the model (4.2), the conditional distribution of $H(Y_{ij})$ with $j \in r_i$ given sampled units is given by

$$H(Y_{ij}) | (y_{ij}, j \in s_i) \sim N(\mathbf{x}_{ij}^t \boldsymbol{\beta} + \tilde{v}_i, \sigma^2), \quad j \in r_i,$$

where $\tilde{v}_i$ is the same as (2.6) except for replacing y_{ij} with $H(y_{ij})$. The best predictor of μ_i is the conditional expectation

$$E[\mu_i | y_{ij}, j \in s_i] = \frac{1}{N_i} \left\{ \sum_{j \in s_i} T(y_{ij}) + \sum_{j \in r_i} E[T(Y_{ij}) | y_{ij}, j \in s_i] \right\}.$$

For general functions $T(\cdot)$ and $H(\cdot)$, the conditional expectation of $T(Y_{ij})$ cannot be obtained in an analytical form, but it can be computed via the Monte Carlo integration by generating random samples of Y_{ij} , where Y_{ij} can be easily simulated via $H^{-1}(U_{ij})$ with U_{ij} generated from $N(\mathbf{x}_{ij}^t \boldsymbol{\beta} + \tilde{v}_i, \sigma^2)$. It should be noted that under the general function $T(\cdot)$, all the information of $\mathbf{x}_{ij}$ is required for computing the predictor of μ_i whereas only the true mean vector of $\mathbf{x}_{ij}$ is required in estimating the mean

parameter as considered in Sect. 2.2. Molina et al. (2014) adopted the hierarchical Bayes approach and developed a Monte Carlo sampling algorithm for estimating μ_i as well as model parameters whereas Molina and Rao (2010) considered the frequentist approach.

4.3 Parametric transformation for response variables

In practice, we often deal with positive response variables such as income, for which the normality assumption used in the FH and NER models could not be reasonable. To address this issue, the log-transformation is widely adopted due to simplicity, and many theoretical results have been revealed (e.g. Slud and Maiti 2006; Molina and Martin 2018). However, the use of log-transformation is not necessarily reasonable, and it would be more preferable to use a parametric family of transformations and estimate the transformation parameter based on the data. Under the FH model, Sugasawa and Kubokawa (2015, 2017c) proposed the following transformed mixed model:

$$H(y_i; \lambda) = \mathbf{x}_i^t \boldsymbol{\beta} + v_i + \varepsilon_i, \quad i = 1, \dots, m,$$

where $H(y_i; \lambda)$ is a parametric transformation, λ is a transformation parameter and the other quantities are defined in the same way as the FH model. Under the model, the areal parameter θ_i is defined as $\theta_i = E[y_i | v_i] = E[H^{-1}(\mathbf{x}_i^t \boldsymbol{\beta} + v_i + \varepsilon_i) | v_i]$. The most famous parametric transformation for positive response variables would be the Box–Cox transformation (Box and Cox 1964), but its inverse function cannot be defined on the whole real line, thereby we cannot define θ_i with the Box–Cox transformation. A possible alternative choice for positive response could be the dual power transformation (Yang 2006) which includes log-transformation as a special case, and this transformation is adopted in Sugasawa and Kubokawa (2015, 2017c). Sugasawa and Kubokawa (2015, 2017c) provided the methods for estimating unknown parameters including λ , the derivation of empirical predictors of θ_i and the asymptotic evaluation of mean squared errors of the predictors.

The same idea can be incorporated into the NER model. Sugasawa and Kubokawa (2019) proposed the following transformed nested error regression model:

$$H(y_{ij}; \lambda) = \mathbf{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, m, \quad (4.3)$$

where $H(y_{ij}; \lambda)$ denotes transformed response variables, and the other quantities are defined in the same way as the NER model. An important application of the transformed NER model is the estimation of general parameters in a finite population as considered in Sect. 4.2. Under the transformed NER model, the method by Molina and Rao (2010) can be easily modified by replacing a fixed transformation with the parametric transformation $H(y_{ij}; \lambda)$. The log-likelihood function without irrelevant constants is given by

$$\begin{aligned}
& -\frac{1}{2} \sum_{i=1}^m \log |\Sigma_i| - \frac{1}{2} \sum_{i=1}^m \{H(\mathbf{y}_i; \lambda) - \mathbf{X}_i \boldsymbol{\beta}\}' \Sigma_i^{-1} \{H(\mathbf{y}_i; \lambda) - \mathbf{X}_i \boldsymbol{\beta}\} \\
& + \sum_{i=1}^m \sum_{j=1}^{n_i} \log \frac{\partial}{\partial \lambda} H(y_{ij}; \lambda),
\end{aligned}$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^t$, $H(\mathbf{y}_i; \lambda) = (H(y_{i1}; \lambda), \dots, H(y_{in_i}; \lambda))^t$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})^t$ and Σ is the variance–covariance matrix of $\mathbf{y}_i$. Under given λ , the maximization with respect to the other parameters is equivalent to the maximum likelihood estimator of the NER model with response variable $H(y_{ij}, \lambda)$, thereby the profile likelihood for λ can be easily obtained. Sugawara and Kubokawa (2019) adopted the golden section method for maximizing the profile likelihood of λ . On the other hand, Sugawara (2019b) developed a hierarchical Bayesian approach similar to Molina et al. (2014) for the model (4.3) with spatially correlated random effects v_i .

4.4 Flexible modeling for variance parts

As mentioned in Sect. 2.1, the main criticism for the FH model (2.1) is the assumption that the sampling variances D_i are known although they are estimated from data. It has been revealed that the assumption for D_i may lead to several serious problems such as underestimation of risks (e.g. Wang and Fuller 2003). To take account of variability of D_i , You and Chapman (2006) introduced the following joint hierarchical model for y_i and D_i :

$$\begin{aligned}
y_i | \theta_i, \sigma_i^2 & \sim N(\theta_i, \sigma_i^2), \quad \theta_i \sim N(\mathbf{x}_i^t \boldsymbol{\beta}, A), \\
D_i | \sigma_i^2 & \sim \text{Ga}\left(\frac{n_i - 1}{2}, \frac{n_i - 1}{2\sigma_i^2}\right), \quad \sigma_i^2 \sim \pi(\sigma_i^2),
\end{aligned} \tag{4.4}$$

where n_i is a sample size, and $\pi(\cdot)$ is a prior distribution for σ_i^2 . In the model (4.4), θ_i and σ_i^2 are the true mean and variance, and y_i and D_i are estimates of them, respectively. You and Chapman (2006) adopted $\text{Ga}(a_i, b_i)$ for $\pi(\cdot)$, where a_i and b_i are fixed constants, and proposed hierarchical Bayesian approach via a Markov Chain Monte Carlo. However, You and Chapman (2006) recommended to use small values for a_i and b_i , thereby the prior for σ_i^2 is diffuse. Under the setting, the resulting estimator of σ_i^2 does not hold shrinkage effect such that the posterior mean of σ_i^2 is almost the same as D_i . To overcome the drawback, Sugawara et al. (2017) proposed the structure $\sigma_i^2 \sim \text{Ga}(a_i, b_i \gamma)$ with unknown γ , and developed an efficient posterior computation algorithm. On the other hand, Dass et al. (2012); Maiti et al. (2014) adopted the model $\sigma_i^2 \sim \text{Ga}(\alpha, \gamma)$, where α and γ are unknown parameters, and developed an empirical Bayes approach.

In the NER model, the variance parameters in random effects and error terms are assumed to be constants in the classical NER model (2.5), but it could be unrealistic for some data as pointed out by Jiang and Nguyen (2012). To address this issue, Jiang and Nguyen (2012) replaced the distribution of v_i and ϵ_{ij} in (2.5) as

$$v_i \sim N(0, \lambda \sigma_i^2), \quad \varepsilon_{ij} \sim N(0, \sigma_i^2), \quad (4.5)$$

where $\lambda, \sigma_1^2, \dots, \sigma_m^2$ are unknown parameters. Since the number of σ_i^2 increases as the number of areas, so that σ_i^2 cannot be consistently estimated as long as n_i is finite. However, an important observation made by the authors is that the predictor of v_i of the form (2.6) depends on the two variance parameters τ^2 and σ^2 only through their ratio τ^2/σ^2 . Under (4.5), the ratio is λ which can be consistently estimated, thereby the predictor of v_i or areal mean can be reasonably derived. The authors also pointed out that the MSE cannot be consistently estimated since it depends on σ_i^2 . To improve the performance of the model (4.5), Kubokawa et al. (2016) introduced additional structure for σ_i^2 , that is, $\sigma_i^2 \sim \text{Ga}(\alpha, \gamma)$, with unknown parameters α and γ . The authors developed likelihood inference on unknown parameters and proved consistency of the maximum likelihood estimator. On the other hand, Sugawara and Kubokawa (2017b) proposed alternative modeling strategy to address heteroscedasticity given by $\text{Var}(v_i) = \tau^2$ and $\text{Var}(\varepsilon_{ij}) = \psi(\mathbf{z}_{ij}^t \boldsymbol{\gamma})$, where $\mathbf{z}_{ij}$ is a sub-vector of $\mathbf{x}_{ij}$, $\boldsymbol{\gamma}$ is a vector of unknown parameters and $\psi(\cdot)$ is a known positive-valued function such as $\exp(\cdot)$. The authors developed moment-based methods to estimate the unknown parameters and derived predictors as well as approximation of MSE.

4.5 Flexible modeling for random effects

As discussed in Sect. 4.1, the random effects play important roles in small area estimation. However, when the true structure does not include random effects, the use of random effects models would increase the estimation variability, so that inclusion of random effects in the model should be carefully examined. To address this issue, Datta et al. (2011) and Molina et al. (2015) proposed testing the existence of random effects before applying mixed models. Datta and Mandal (2015) generalized this idea to the mixture modeling for random effects in the FH model (2.1), which is given by

$$v_i | (s_i = 1) \sim N(0, A), \quad v_i | (s_i = 0) \sim \delta_0, \quad P(s_i = 1) = \pi, \quad (4.6)$$

where δ_0 is the one-point distribution on the origin, and s_i is a latent random variable indicating whether random area effect should be needed or not. Then the unknown parameter π in the model (4.6) can be interpreted as the prior probability of existence of random effects in the i th area. Datta et al. (2011) called the random effects structure (4.6) Muncertain random effects'. It is noted that the marginal distribution of v_i is the mixture of two distributions $\pi N(0, A) + (1 - \pi)\delta_0$, which is very similar to the spike-and-slab prior (e.g. Ishwaran and Rao 2005) for variable selection. Datta et al. (2011) proposed a hierarchical Bayesian approach to estimate θ_i . Sugawara and Kubokawa (2017a) adopted the idea (4.6) in the NER model to demonstrate the usefulness in the context of estimating parameters in a finite population framework. On the other hand, Sugawara et al. (2017) generalized the idea of the uncertain random effects (4.6) to the area-level model based on the natural exponential family (2.8). The new model for θ_i is given by

$$\theta_i | (s_i = 1) \sim \pi(\theta_i), \quad \theta_i | (s_i = 0) = (\psi')^{-1}(m_i), \quad P(s_i = 1) = \pi,$$

where m_i is the mean of the conjugate prior and $(\psi')^{-1}$ denotes the inverse function of ψ' given in Sect. 2.5. Based on the model, Sugawara et al. (2017) developed an empirical Bayes method for estimating $\mu_i = E[y_i | \theta_i]$. As a related method, Chakraborty et al. (2016) proposed a two component mixture model $\pi N(0, A_1) + (1 - \pi)N(0, A_2)$ with unknown A_1 and A_2 for v_i .

A new direction for modeling random effects is the use of the global–local shrinkage prior originally developed in the context of signal estimation (e.g. Carvalho et al. 2010). Tang et al. (2018) introduced the global–local shrinkage prior for molding random effects in the FH model (2.1). Their model is given by

$$v_i | \lambda_i^2, \tau^2 \sim N(0, \lambda_i^2 \tau^2), \quad \lambda_i^2 \sim \pi(\lambda_i^2).$$

For example, the exponential prior for λ_i^2 leads to the Laplace prior for v_i . Under the formulation, the conditional expectation of θ_i is $y_i - B_{GL,i}(y_i - \mathbf{x}_i^t \boldsymbol{\beta})$ for $B_{GL,i} = D_i / (D_i + \lambda_i^2 \tau^2)$. Tang et al. (2018) showed that the shrinkage coefficient $B_{GL,i}$ converges to 0 as $|y_i - \mathbf{x}_i^t \boldsymbol{\beta}| \rightarrow \infty$, which means that the resulting estimator can prevent over-shrinkage issues, that is, it does not shrink the direct estimator which is very far from the regression part. Under count responses, Hamura et al. (2019) developed a new global–local shrinkage prior under the Poisson–gamma model given in Sect. 2.5.

4.6 Skewed distributions for error terms and random effects

In the classical FH and NER model, the distributions of both error terms and random effects are assumed to be normal, which are not necessarily reasonable in practice. In the FH model (2.1), the normality assumption of the error term ε_i is based on central limit theorems since y_i is typically a summary statistic. However, when the sample size for computing y_i is small, the normality assumption would be violated. Especially, empirical distributions of data such as income tend to be skewed. Then Ferraz and Mourab (2012) adopted the skew-normal distribution (e.g. Azzalini 1985) for ε_i in the FH model. Ferrante and Pacei (2017) developed a multivariate FH model with both error terms and random effects following multivariate skew-normal distributions. In the context of the NER models, Diallo and Rao (2018) was concerned with the situation where the distributions of the response variables are skewed. The authors proposed replacing the normality assumption, namely $v_i \sim N(0, \tau^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$ with $v_i \sim \text{SN}(-\delta_v \tau \sqrt{2/\pi}, \tau^2, \lambda_v)$ and $\varepsilon_{ij} \sim \text{SN}(-\delta_\varepsilon \sigma \sqrt{2/\pi}, \sigma^2, \lambda_\varepsilon)$, where $\delta_v = \lambda_v / \sqrt{1 + \lambda_v^2}$, $\delta_\varepsilon = \lambda_\varepsilon / \sqrt{1 + \lambda_\varepsilon^2}$ and $\text{SN}(\mu, \sigma^2, \lambda)$ is a skew-normal distribution with density

$$f(x; \mu, \sigma^2, \lambda) = \frac{2}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right) \Phi\left(\lambda \frac{x - \mu}{\sigma}\right)$$

with $\phi(\cdot)$ and $\Phi(\cdot)$ being the density and distribution functions of the standard normal distribution. Note that the non-zero location parameters in v_i and ε_{ij} ensure that

$E[v_i] = 0$ and $E[\varepsilon_{ij}] = 0$. However, it may be hard to estimate τ^2 , σ^2 , λ_v and λ_ε accurately. Tsujino and Kubokawa (2019) investigated a simplified version of the model in which only error terms follow the skew-normal distribution and the random effects are still normal, and provided a simple expression for the predicting random effects.

4.7 Nonparametric and semiparametric modeling

In the standard mixed models in small area estimation, parametric models are typically adopted for simplicity. However, parametric models suffer from model misspecification, which could produce unreliable small area estimates. Instead of using parametric models, the use of nonparametric or semiparametric models have been considered in the literature.

Opsomer et al. (2008) considered nonparametric estimation of the regression part in the linear mixed model by adopting the P-spline method. Specifically, the authors proposed the following model:

$$y_{ij} = f(x_{ij}) + v_i + \varepsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, m, \quad (4.7)$$

where v_i and ε_{ij} are defined in the same way as the original NER model (2.5), and $f(\cdot)$ is modeled by the P-spline given by

$$f(x) = \beta_0 + \beta_1 x + \dots + \beta_q x^q + \sum_{\ell=1}^K \gamma_\ell (x - \kappa_\ell)_+^q,$$

where q is the degree of the spline, $(x)_+^q = x^q I(x > 0)$, $\kappa_1 < \dots < \kappa_K$ is a set of fixed knots and $\beta = (\beta_0, \dots, \beta_q)$ and $\gamma = (\gamma_1, \dots, \gamma_K)$ denote coefficient vectors for the parametric and spline terms. Moreover, it is assumed that $\gamma \sim N(0, \lambda I_K)$, where I_K is the $K \times K$ identity matrix, and λ is an unknown parameter controlling the smoothness of the estimation of $f(\cdot)$. Let $\mathbf{x}_{ij} = (1, x_{ij}, \dots, x_{ij}^q)$ and $\mathbf{w}_{ij} = ((x - \kappa_1)_+^q, \dots, (x - \kappa_K)_+^q)$. Then the nonparametric model (4.7) can be expressed as the linear mixed model

$$y_{ij} = \mathbf{x}_{ij}^t \beta + \mathbf{w}_{ij}^t \gamma + v_i + \varepsilon_{ij},$$

and the unknown parameters are estimated via the likelihood method. The P-spline is also adopted for modeling regression terms in the literature (e.g. Rao et al. 2014; Sugawara et al. 2018).

Sugawara et al. (2019) is concerned with misspecification of the link function in the unmatched sampling and linking model (2.7). Although the link function is typically specified by a user, the selected link function is subject to misspecification. To overcome the problem, Sugawara et al. (2019) employed the P-spline technique to estimate the link function and small area parameters simultaneously. The proposed model is given by

$$y_i | u_i, \gamma_2 \sim N(z_1(u_i)' \gamma_1 + z_2(u_i; \delta)' \gamma_2, D_i),$$

$$u_i \sim N(x_i' \beta, A), \quad \gamma_2 \sim N(0, \lambda I_K),$$

where $z_1(u_i) = (1, u_i, \dots, u_i^q)'$, $\gamma_1 = (\gamma_{10}, \dots, \gamma_{1q})'$ and $z_2(u_i; \delta) = ((u_i - \kappa_1)_+^q, \dots, (u_i - \kappa_K)_+^q)'$. Here, $\theta_i = z_1(u_i)' \gamma_1 + z_2(u_i; \delta)' \gamma_2$ is the small area parameter. The authors put prior distributions on the unknown parameters and developed a hierarchical Bayesian method for estimating θ_i .

Finally, to avoid specification of the distributions of random effects or error terms, Chaudhuri and Ghosh (2011) adopted an empirical likelihood approach for both area-level and unit-level models, and Polettini (2017) employed Dirichlet process mixture modeling for random effects in the FH model.

4.8 Robust methods

In practice, outliers are often contained in data. The model assumption such as normality would be violated for such outlying data, and the inclusion of outliers would significantly affect estimation of model parameters. In the context of the FH model, Ghosh et al. (2008) used the influence function to investigate the effect of outliers on the BLUP, and proposed the robust version of BLUP given by

$$y_i - \frac{D_i v_i(A)^{1/2}}{A + D_i} \psi_K \left(\frac{y_i - x_i' \hat{\beta}_{\text{GLS}}}{v_i(A)^{1/2}} \right) \quad \text{with} \quad v_i(A) = A + D_i - x_i' \left(\sum_{i=1}^m \frac{x_i x_i'}{A + D_i} \right)^{-1} x_i,$$

where $\psi_K(t) = u \min(1, K/|u|)$ is Huber's ψ -function with a tuning constant $K > 0$. Similarly, Sinha and Rao (2009) used Huber's ψ -function to modify an equation for θ_i and suggested a robust predictor as a solution to the equation

$$D_i^{-1/2} \psi_K \left\{ D_i^{-1/2} (y_i - \theta_i) \right\} - A^{-1/2} \psi_K \left\{ A^{-1/2} (\theta_i - x_i' \beta) \right\} = 0.$$

Moreover, Sugawara (2019a) proposed the use of density power divergence given by

$$L_\alpha(\beta, A) = \frac{1}{\alpha} \sum_{i=1}^m \phi(y_i; x_i' \beta, A + D_i)^\alpha - \{2\pi(A + D_i)\}^{-\alpha/2} (1 + \alpha)^{-3/2},$$

where $\phi(\cdot; a, b)$ denotes the density function of the normal distribution with mean a and variance b , and α is a tuning parameter. Note that $L_\alpha(\beta, A)$ reduces to the standard marginal likelihood function without an irrelevant constant as $\alpha \rightarrow 0$. Using the fact that the Bayes predictor $\tilde{\theta}_i$ under the FH model (2.1) can be derived via the partial derivatives of the marginal likelihood with respect to y_i , Sugawara (2019a) defined a new robust estimator by replacing the marginal likelihood with the density power divergence, and the robust predictor is given by

$$\tilde{\theta}_i^R = y_i - \frac{D_i}{A + D_i} (y_i - x_i' \beta) \phi(y_i; x_i' \beta, A + D_i)^\alpha.$$

It should be noted that $\tilde{\theta}_i^R$ reduces to the original predictor $\tilde{\theta}_i$ as $\alpha \rightarrow 0$. Compared with $\tilde{\theta}_i$, the shrinkage coefficient depends on y_i through its density function. Since the second term converges to 0 as $|y_i| \rightarrow \infty$ under fixed parameters, the estimator $\tilde{\theta}_i^R$ reduces to y_i for outlying observation y_i .

On the other hand, Datta and Lahiri (1995) replaced the normal distribution for the random effects with the Cauchy distribution in outlying areas, and provided some robustness properties of the resulting estimator. Moreover, Ghosh et al. (2018) studied the use of the t -distribution for the random effects under the FH model with joint modeling means and variances as considered in Sugasawa et al. (2017).

In estimating population parameters in the framework of finite population, existing outliers in the sampled data can invalidate the parametric assumptions of the NER model (2.5), which might impact the validity of model-based small area estimators. To solve the problem, we can again employ the idea given in Sinha and Rao (2009) to get robust estimator $\hat{\beta}^R$ and $\hat{v}_i^R$, which enables us to obtain robust small area estimators of the mean in the i th area, given by

$$\frac{1}{N_i} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} (x_{ij}^t \hat{\beta}^R + \hat{v}_i^R) \right\},$$

where s_i and r_i are the sets of indices of sampled and non-sampled units, respectively. However, the above estimator assumes that the non-sampled values in the population are drawn from a distribution with the same mean as the sampled non-outliers, which can be unrealistic. Dongmo-Jiongo et al. (2013) and Chambers et al. (2014) addressed this issue and proposed an improved version of the above robust estimator. Moreover, Schmid et al. (2016) extended the results of Chambers et al. (2014) to take spatial correlation into account.

4.9 Measurement error models

In the FH model (2.1), covariate or auxiliary information x_i is fixed. However, x_i could be estimates from another survey in practice, thereby x_i might include estimation error, that is, x_i is measured with errors. Ybarra and Lohr (2008) clearly pointed out the drawback of simply using x_i under measurement error by investigating MSE. Let $\hat{x}_i$ be the estimator of the true covariate x_i , and suppose that its MSE matrix $\text{MSE}(\hat{x}_i) = C_i$ is known. Note that the true areal parameter is $\theta_i = x_i^t \beta + v_i$, and the naive predictor with $\hat{x}_i$ and unknown parameters is $\tilde{\theta}_i^N = \gamma_i y_i + (1 - \gamma_i) \hat{x}_i^t \beta$. Ybarra and Lohr (2008) derived the MSE of $\tilde{\theta}_i^N$ as

$$\text{MSE}(\tilde{\theta}_i^N) = \text{MSE}(\tilde{\theta}_i) + (1 - \gamma_i) \beta^t C_i \beta = D_i \gamma_i + (1 - \gamma_i) \beta^t C_i \beta,$$

where $\tilde{\theta}_i$ is the predictor in the known case of x_i . This suggests that the measurement error increases MSE of the standard predictor. More importantly, since the MSE of the direct estimator y_i is D_i , the naive predictor $\tilde{\theta}_i$ could be worse than the naive estimator if $\beta^t C_i \beta > A + D_i$. Ybarra and Lohr (2008) derived an optimal predictor under the measurement error, which is obtained by replacing γ_i with $\tilde{\gamma}_i = (A + \beta^t C_i \beta) / (A + D_i + \beta^t C_i \beta)$. The unknown parameters can be estimated via

the maximum likelihood method, which enables us to get the empirical version of the predictor. On the other hand, Arima et al. (2015) pointed out that the parameter estimation could be unstable, and the authors developed a hierarchical Bayesian approach for the models by Ybarra and Lohr (2008).

Ghosh et al. (2006) and Torabi et al. (2009) studied the NER model (2.5) with measurement errors. The authors assume that the true covariate is x_i for all the units in the same area, and different measurements x_{ij} are observed for different sampled units. It is assumed that $x_{ij} = x_i + \eta_{ij}$, where $x_i \sim N(\mu_x, \sigma_x^2)$ and $\eta_{ij} \sim N(0, \gamma^2)$. This formulation is called structural measurement error. Note that μ_x , σ_x^2 and γ^2 are additional unknown parameters, which can be estimated via the maximum likelihood method.

4.10 Observed best prediction

The classical FH model (2.1) implicitly assumes that the regression part $\mathbf{x}_i^t \boldsymbol{\beta}$ is correctly specified, and the estimation of model parameters including $\boldsymbol{\beta}$ as well as the prediction of θ_i are carried out under the assumed model. However, any assumed model is subject to model misspecification. Jiang et al. (2011) introduced the true model $\theta_i = \mu_i + v_i$ with $v_i \sim N(0, A)$, where μ_i is the true mean which is not necessarily equivalent to $\mathbf{x}_i^t \boldsymbol{\beta}$. Note that the assumed model (2.1) does not change, and they focused on a reasonable estimation method for regression coefficients $\boldsymbol{\beta}$ under possible model misspecification. They considered the total mean squared prediction error (MSPE) of the best predictor $\tilde{\theta}_i$ given by

$$\text{MSPE}(\boldsymbol{\beta}) = \sum_{i=1}^m \{B_i y_i - \theta_i + \mathbf{x}_i^t \boldsymbol{\beta} (1 - B_i)\}^2,$$

where $B_i = A/(A + D_i)$, and an unbiased estimator of the MSPE is

$$\text{MSPE}(\boldsymbol{\beta}) = E \left[\sum_{i=1}^m (B_i y_i - \theta_i)^2 + \sum_{i=1}^m (1 - B_i)^2 (\mathbf{x}_i^t \boldsymbol{\beta})^2 - 2 \sum_{i=1}^m (1 - B_i)^2 \mathbf{x}_i^t \boldsymbol{\beta} y_i \right].$$

This expression suggests a natural estimator of $\boldsymbol{\beta}$ as the minimizer of the expression inside the expectation, which is equivalent to minimizing

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^m (1 - B_i)^2 (\mathbf{x}_i^t \boldsymbol{\beta})^2 - 2 \sum_{i=1}^m (1 - B_i)^2 \mathbf{x}_i^t \boldsymbol{\beta} y_i.$$

Then a closed form expression of the minimizer is obtained as

$$\hat{\boldsymbol{\beta}}_{\text{OBP}} = \left\{ \sum_{i=1}^m (1 - B_i)^2 \mathbf{x}_i \mathbf{x}_i^t \right\}^{-1} \sum_{i=1}^m (1 - B_i)^2 \mathbf{x}_i y_i, \quad (4.8)$$

which is called the observed best predictive (OBP) estimator of $\boldsymbol{\beta}$. Note that the expression (4.8) is different from the maximum likelihood estimator or GLS estimator given in (2.3). In Jiang et al. (2011), they also developed general OBP estimators

under linear mixed models and asymptotic theory of OBP estimators. As follow-up works, Jiang et al. (2011) is concerned with misspecification of both mean and variance parts in the NER model (2.5), and Chen et al. (2015) extended the OBP theory to a situation with binary responses.

4.11 Variable selection

Modeling of regression parts is an important issue to obtain a good small area estimator, thereby variable selection problems would arise naturally. In the general statistical theory, variable selection has been recognized as essential problems since the proposal of well-known *Akaike Information Criterion* or AIC (Akaike 1973, 1974). *Bayesian Information Criterion* or BIC (Schwarz 1978) is also a well-known criterion, but the use of both AIC and BIC is not necessarily reasonable for variable selection when the model contains random effects.

Vaida and Blanchard (2005) considered the general form of linear mixed models given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \quad \mathbf{v} \sim \mathbf{N}(\mathbf{0}, \mathbf{G}), \quad \mathbf{e} \sim \mathbf{N}(\mathbf{0}, \mathbf{R}),$$

where $\mathbf{X}$ and $\mathbf{Z}$ are design matrices for fixed effects and random effects, respectively. Vaida and Blanchard (2005) pointed out that the classical AIC measures the prediction error of the predictor based on the marginal distribution $\mathbf{y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, thereby the classical AIC might be inappropriate for predicting quantity including random effects. Therefore, the authors proposed the conditional AIC as an asymptotically unbiased estimator of the conditional Akaike information (cAI) defined by

$$\text{cAI} = -2 \int E_{\mathbf{y}, \mathbf{y}^* | \mathbf{v}} \left[\log \phi(\mathbf{y}^*; \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{v}}, \hat{\mathbf{R}}) \right] \phi(\mathbf{v}; \mathbf{0}, \mathbf{G}) d\mathbf{v}, \quad (4.9)$$

where $E_{\mathbf{y}, \mathbf{y}^* | \mathbf{v}}$ denotes the expectation with respect to the conditional distribution of $\mathbf{y}$ and $\mathbf{y}^*$ given $\mathbf{v}$, namely, $\mathbf{y} | \mathbf{v} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \mathbf{R})$ and $\mathbf{y}^* | \mathbf{v} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \mathbf{R})$, respectively. Note that cAI is related to the Kullback–Leibler divergence between the estimated and true conditional density of a future observation $\mathbf{y}^*$. Vaida and Blanchard (2005) derived an asymptotically unbiased estimator of cAI of the form $-2 \log \phi(\mathbf{y}^*; \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{v}}, \hat{\mathbf{R}}) - \Delta$ with bias correction term Δ , which is called conditional AIC. Since both FH and NER models are special cases of linear mixed models, the cAIC can be adopted for selecting suitable covariates. We refer to Muller et al. (2013) for more details and review of variable selection techniques in linear mixed models.

One possible criticism in the deviation of cAIC is that the same design matrix $\mathbf{X}$ is used for both estimation and prediction models. However, in the context of the estimating population parameters via the NER model as described in Sect. 2.2, the two models have different covariates since we are interested in the prediction based on $\bar{\mathbf{X}}_i$ and the NER model is based on $\{\mathbf{x}_{ij}\}_{j=1, \dots, n_i, i=1, \dots, m}$. Kawakubo et al. (2018) addressed this issue and developed a new information criterion for variable selection under this situation by modifying the derivation of cAIC. We also note that

minimizing cAI is not necessarily equivalent to minimizing the mean squared error of small area estimators. Hence, if we pursue suitable variable selection to obtain better predictors which have smaller mean squared error, it would be more reasonable to derive an information criterion based on mean squared errors. Sugawara et al. (2019) adopted the OBP method given in Sect. 4.10 to derive a criterion for conducting variable selection based on unbiased estimator of mean squared prediction errors.

As a more general strategy, Jiang et al. (2008) proposed a variable selection procedure called the fence method in generalized linear mixed models, and applied the method for selecting covariates in the FH and NER models. Moreover, Jiang et al. (2010) adopted the fence method for fitting the P-spline models described in Sect. 4.7, which requires selecting the degree of the spline, the number of knots, and a smoothing parameter.

Finally, it should be noted that the variable selection and estimation are typically considered separately, which might lead to ignoring variability in the selection procedure. This issue is addressed by Jiang et al. (2018) and the authors introduced a new jackknife procedure for estimating mean squared errors that takes accounts of additional variability due to variable selection.

4.12 Other related topics

We here briefly review some related topics that we did not cover the previous section. One of them is the benchmarking. This is a method for adjusting the EBLUP so that the total sum of EBLUPs is equal to that of the direct estimates, and several new methodologies have been recently developed (e.g. Kubokawa and Strawderman 2013; Kubokawa et al. 2014; Ghosh et al. 2015). There have been several attempts to propose a new methodology for small area estimation of non-standard data such as grouped data (Kawakubo and Kobayashi 2019) and circular data (Hernandez-Stumpfhauser et al. 2016). Torabi and Rao (2013) proposed an extension of the FH model under existing sub-areas. Finally, the use of spatial and time series data is getting more and more important to produce reliable estimates by borrowing information from both spatially and temporally close areas. Some extensions of the standard mixed models such as FH and NER models have been considered in the literature (e.g. Pratesi and Salvati 2008, 2009; Marhuenda et al. 2013).

5 Concluding remarks

In this paper, we reviewed both classical and recently developed methods for small area estimation with mixed models. Since this paper has even focused on small area estimation based on mixed models, we did not cover other methodologies without using mixed models or more practical issues such as survey weights. Please see Rao and Molina (2015) for details of such approaches.

Acknowledgements The authors are grateful to the Editor, the Associate Editor and two reviewers for their valuable comments and helpful suggestions. Research of the authors was supported in part by

Grant-in-Aid for Scientific Research (18K11188, 18K12757) from Japan Society for the Promotion of Science.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akademia Kiado.
- Akaike, H. (1974). A new look at the statistical model identification. *System identification and time-series analysis. IEEE Transaction on Automatic Control AC*, 19, 716–723.
- Arima, S., Datta, G. S., & Liseo, B. (2015). Bayesian estimators for small area models when auxiliary information is measured with error. *Scandinavian Journal of Statistics*, 42, 518–529.
- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12, 171–178.
- Basu, R., Ghosh, J., & Mukerjee, R. (2003). Empirical bayes prediction intervals in a normal regression model: Higher order asymptotics. *Statistics and Probability Letters*, 63, 197–203.
- Battese, G., Harter, R., & Fuller, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28–36.
- Booth, J. S., & Hobert, P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 262–272.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformation (with discussion). *Journal of the Royal Statistical Society: Series B*, 26, 211–252.
- Butar, F. B., & Lahiri, P. (2003). On measures of uncertainty of empirical bayes small-area estimators. *Journal of Statistical Planning and Inference*, 112, 63–76.
- Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97, 465–480.
- Chakraborty, A., Datta, G. S., & Mandal, A. (2016). A two-component normal mixture alternative to the Fay–Herriot model. *Statistics in Transition New Series*, 17, 67–90.
- Chambers, R., Chandra, H., Salvati, N., & Tzavidis, N. (2014). Outliner robust small area estimation. *Journal of the Royal Statistical Society: Series B*, 76, 47–69.
- Chatterjee, S., Lahiri, P., & Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related predictions intervals in linear mixed models. *The Annals of Statistics*, 36, 1221–1245.
- Chaudhuri, S., & Ghosh, M. (2011). Empirical likelihood for small area estimation. *Biometrika*, 98, 473–480.
- Chen, S., Jiang, J., & Nguyen, T. (2015). Observed best prediction for small area counts. *Journal of Survey Statistics and Methodology*, 3, 136–161.
- Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671–681.
- Das, K., Jiang, J., & Rao, J. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 818–840.
- Dass, S. C., Maiti, T., Ren, H., & Sinha, S. (2012). Confidence interval estimation of small area parameters shrinking both means and variances. *Survey Methodology*, 38, 173–187.
- Datta, G., Rao, J., & Smith, D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, 92, 183–196.

- Datta, G. S., Ghosh, M., Smith, D. D., & Lahiri, P. (2002). On an asymptotic theory of conditional and unconditional coverage probabilities of empirical bayes confidence intervals. *Scandinavian Journal of Statistics*, 29, 139–152.
- Datta, G. S., Hall, P., & Mandal, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, 106, 362–374.
- Datta, G. S., & 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., & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613–627.
- Datta, G. S., & Mandal, A. (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, 110, 1735–1744.
- Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New York: Wiley.
- Diallo, M. S., & Rao, J. N. K. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, 45, 1092–1116.
- Diao, L., Smith, D. D., Datta, G. S., Maiti, T., & Opsomer, J. D. (2014). Accurate confidence interval estimation of small area parameters under the Fay–Herriot model. *Scandinavian Journal of Statistics*, 41, 497–515.
- Dongmo-Jiongo, V., Haziza, D., & Duchesne, P. (2013). Controlling the bias of robust small area estimators. *Biometrika*, 100, 843–858.
- Fay, R., & Herriot, R. (1979). Estimators of income for small area places: An application of James–Stein procedures to census. *Journal of the American Statistical Association*, 74, 341–353.
- Ferrante, M. R., & Pacei, S. (2017). Small domain estimation of business statistics by using multivariate skew normal models. *Journal of the Royal Statistical Society: Series A*, 180, 1057–1088.
- Ferraz, V., & Mourab, F. A. S. (2012). Small area estimation using skew normal models. *Computational Statistics and Data Analysis*, 56, 2864–2874.
- Foster, J., Greer, J., & Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, 52, 761–766.
- Ghosh, M., Kubokawa, T., & Kawakubo, Y. (2015). Benchmarked empirical bayes methods in multiplicative area-level models with risk evaluation. *Biometrika*, 102, 647–659.
- Ghosh, M., & Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*, 91, 95–112.
- Ghosh, M., Maiti, T., & Roy, A. (2008). Influence functions and robust Bayes and empirical Bayes small area estimation. *Biometrika*, 95, 573–585.
- Ghosh, M., Myung, J., & Moura, F. A. S. (2018). Robust Bayesian small area estimation. *Survey Methodology* 44(12), 001–X.
- Ghosh, M., Natarajan, K., Stroud, T. W. F., & Carlin, B. P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273–282.
- Ghosh, M., & Rao, J. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55–93.
- Ghosh, M., Sinha, K., & Kim, D. (2006). Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error models. *Scandinavian Journal of Statistics*, 33, 591–608.
- Hall, P., & Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics*, 34, 1733–1750.
- Hall, P., & Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B*, 68, 221–238.
- Hamura, H., K. Irie, & S. Sugawara. (2019). On global-local shrinkage priors for count data. arXiv:1907.01333.
- Henderson, C. (1950). Estimation of genetic parameters. *The Annals of Mathematical Statistics*, 21, 309–310.
- Hernandez-Stumpfhauser, D., Breidt, J. F., & Opsomer, J. D. (2016). Hierarchical bayesian small area estimation for circular data. *Canadian Journal of Statistics*, 44, 416–430.
- Hirose, M. (2017). Non-area-specific adjustment factor for second-order efficient empirical Bayes confidence interval. *Computational Statistics and Data Analysis*, 116, 67–78.
- Hirose, M. (2019). A class of general adjusted maximum likelihood methods for desirable mean squared error estimation of EBLUP under the fay-herriot small area model. *Journal of Statistical Planning and Inference*, 199, 302–310.

- Hirose, M., & Lahiri, P. (2018). Estimating variance of random effects to solve multiple problems simultaneously. *The Annals of Statistics*, 46, 1721–1741.
- Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33, 730–773.
- Jiang, J., & Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53, 217–243.
- Jiang, J., & Lahiri, P. (2006). Estimation of finite population domain means: a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301–311.
- Jiang, J., Lahiri, P., & Wan, S. M. (2002). A unified Jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics*, 30, 1782–1810.
- Jiang, J., & Nguyen, T. (2012). Small area estimation via heteroscedastic nested-error regression. *Canadian Journal of Statistics*, 40, 588–603.
- Jiang, J., Nguyen, T., & Lahiri, P. (2018). A unified Monte-Carlo Jackknife for small area estimation after model selection. *Annals of Mathematical Sciences and Applications*, 3, 405–438.
- Jiang, J., Nguyen, T., & Rao, J. S. (2010). Fence method for non-parametric small area estimation. *Survey Methodology*, 36, 3–11.
- Jiang, J., Nguyen, T., & Rao, J. S. (2011). Best predictive small area estimation. *Journal of the American Statistical Association*, 106, 732–745.
- Jiang, J., Rao, J. S., Gu, Z., & Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics*, 36, 1669–1692.
- Kawakubo, Y., & Kobayashi, G. (2019). Small area estimation of general finite-population parameters based on grouped data. arXiv:1903.07239.
- Kawakubo, Y., Sugawara, S., & Kubokawa, T. (2018). Conditional Akaike information under covariate shift with application to small area estimation. *Canadian Journal of Statistics*, 46, 316–335.
- Kubokawa, T. (2009). Corrected empirical bayes confidence intervals in nested error regression models. *Journal of the Korean Statistical Society*, 39, 221–236.
- Kubokawa, T., Hasukawa, M., & Takahashi, K. (2014). On measuring uncertainty of benchmarked predictors with application to disease risk estimate. *Scandinavian Journal of Statistics*, 41, 394–413.
- Kubokawa, T., & Strawderman, W. E. (2013). Dominance properties of constrained Bayes and empirical Bayes estimators. *Bernoulli*, 19, 2220–2221.
- Kubokawa, T., Sugawara, S., Ghosh, M., & Chaudhuri, S. (2016). Prediction in heteroscedastic nested error regression models with random dispersions. *Statistica Sinica*, 26, 465–492.
- Lahiri, P., & Rao, J. N. K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90, 758–766.
- Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882–892.
- Lohr, S. L., & Rao, J. N. K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika*, 96, 457–468.
- Maiti, T., Ren, H., & Sinha, A. (2014). Prediction error of small area predictors shrinking both means and variances. *Scandinavian Journal of Statistics*, 41, 775–790.
- Marhuenda, Y., Molina, I., & Morales, D. (2013). Small area estimation with spatio-temporal Fay–Herriot models. *Computational Statistics and Data Analysis*, 58, 308–325.
- Mauro, F., Monleon, V. J., Temesgen, H., & Ford, K. R. (2017). Analysis of area level and unit level models for small area estimation in forest inventories assisted with Lidar auxiliary information. *PLoS One*, 12, e0189401.
- Molina, I., & Marhuenda, Y. (2015). SAE: An R package for small area estimation. *The R Journal*, 7, 81–98.
- Molina, I., & Martin, N. (2018). Empirical best prediction under a nested error model with log transformation. *The Annals of Statistics*, 46, 1961–1993.
- Molina, I., Nandram, B., & Rao, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical bayes approach. *The Annals of Applied Statistics*, 8, 852–885.
- Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38, 369–385.
- Molina, I., Rao, J. N. K., & Datta, G. S. (2015). Small area estimation under a Fay–Herriot model with preliminary testing for the presence of random area effects. *Survey Methodology*, 41, 1–19.
- Muller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, 28, 135–167.

- Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., & Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B*, 70, 265–286.
- Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28, 40–68.
- Polettini, S. (2017). A generalized semiparametric Bayesian Fay–Herriot model for small area estimation shrinking both means and variances. *Bayesian Analysis*, 12, 729–752.
- Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163–171.
- Pratesi, M. (2016). *Analysis of poverty data by small area estimation*. New York: Wiley.
- Pratesi, M., & Salvati, N. (2008). Small area estimation: The EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, 17, 113–141.
- Pratesi, M., & Salvati, N. (2009). Small area estimation in the presence of correlated random area effects. *Journal of Official Statistics*, 25, 37–53.
- Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd ed.). New York: Wiley.
- Rao, J. N. K., Sinha, S. K., & Dumitrescu, L. (2014). Robust small area estimation under semi-parametric mixed models. *Canadian Journal of Statistics*, 42, 126–141.
- Schmid, T., Bruckschen, F., Salvati, N., & Zbiranski, T. (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society: Series A*, 180, 1163–1190.
- Schmid, T., Tzavidis, N., Münnich, R., & Chambers, R. (2016). Outlier robust small-area estimation under spatial correlation. *Scandinavian Journal of Statistics*, 43, 806–826.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.
- Sinha, S. K., & Rao, J. N. K. (2009). Robust small area estimation. *Canadian Journal of Statistics*, 37, 381–399.
- Slud, E., & Maiti, T. (2006). Mean-squared error estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society: Series B*, 68, 239–257.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 197–206.
- Sugasawa, S. (2019a). Robust empirical Bayes small area estimation with density power divergence. *Biometrika*(to appear).
- Sugasawa, S. (2019b). Small area estimation of general parameters: Bayesian transformed spatial prediction approach. *Japanese Journal of Statistics and Data Science*(to appear).
- Sugasawa, S., Kawakubo, Y., & Datta, G. S. (2019). Observed best selective prediction in small area estimation. *Journal of Multivariate Analysis*, 173, 383–392.
- Sugasawa, S., & Kubokawa, T. (2015). Parametric transformed Fay–Herriot model for small area estimation. *Journal of Multivariate Analysis*, 139, 17–33.
- Sugasawa, S., & Kubokawa, T. (2016). On conditional prediction errors in mixed models with application to small area estimation. *Journal of Multivariate Analysis*, 148, 18–33.
- Sugasawa, S., & Kubokawa, T. (2017a). Bayesian estimators in uncertain nested error regression models. *Journal of Multivariate Analysis*, 153, 52–63.
- Sugasawa, S., & Kubokawa, T. (2017b). Heteroscedastic nested error regression models with variance functions. *Statistica Sinica*, 27, 1101–1123.
- Sugasawa, S., & Kubokawa, T. (2017c). Transforming response values in small area prediction. *Computational Statistics and Data Analysis*, 114, 47–60.
- Sugasawa, S., & Kubokawa, T. (2019). Adaptively transformed mixed model prediction of general finite population parameters. *Scandinavian Journal of Statistics*, 46, 1025–1046.
- Sugasawa, S., Kubokawa, T., & Ogasawara, K. (2017). Empirical uncertain Bayes methods in area-level models. *Scandinavian Journal of Statistics*, 44, 684–706.
- Sugasawa, S., Kubokawa, T., & Rao, J. N. K. (2018). Small area estimation via unmatched sampling and linking models. *TEST*, 27, 407–427.
- Sugasawa, S., Kubokawa, T., & Rao, J. N. K. (2019). Hierarchical Bayes small area estimation with an unknown link function. *Scandinavian Journal of Statistics*, 46, 885–897.
- Sugasawa, S., Tamae, H., & Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44, 150–167.

- Tang, X., Ghosh, M., Ha, N. S., & Sedransk, J. (2018). Modeling random effects using global-local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, 113(524), 1476–1489.
- Torabi, M., Datta, G. S., & Rao, J. N. K. (2009). Empirical Bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates. *Scandinavian Journal of Statistics*, 2009, 355–368.
- Torabi, M., & Rao, J. (2013). Estimation of mean squared error of model-based estimators of small area means under a nested error linear regression model. *Journal of Multivariate Analysis*, 117, 76–87.
- Tsujino, T., & Kubokawa, T. (2019). Empirical Bayes methods in nested error regression models with skew-normal errors. *Japanese Journal of Statistics and Data Science*, 2, 375–403.
- Vaida, F., & Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, 92, 351–370.
- Wang, J., & Fuller, W. (2003). The mean squared error of small area predictors constructed with estimated error variances. *Journal of the American Statistical Association*, 98, 716–723.
- Williams, D. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, 31, 949–952.
- Yang, Z. L. (2006). A modified family of power transformations. *Economics Letters*, 92, 14–19.
- Ybarra, L. M. R., & Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95, 919–931.
- Yoshimori, M., & Lahiri, P. (2014a). A new adjusted maximum likelihood method for the Fay–Herriot small area model. *Journal of Multivariate Analysis*, 124, 281–294.
- Yoshimori, M., & Lahiri, P. (2014b). A second-order efficient empirical bayes confidence interval. *The Annals of Statistics*, 42, 1233–1261.
- You, Y., & Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97–103.
- You, Y., & Rao, J. N. K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, 30, 3–15.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.