Statistica Sinica Preprint No: SS-2023-0045	
Title	Small Area Estimation using EBLUPs under the Nested Error Regression Model
Manuscript ID	SS-2023-0045
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0045
Complete List of Authors	Ziyang Lyu and A.H. Welsh
Corresponding Authors	Ziyang Lyu
E-mails	lvziyang08@gmail.com
Notice: Accepted version subject to English editing.	

Statistica Sinica

Small Area Estimation using EBLUPs under the Nested Error Regression Model

Ziyang Lyu[1] and A.H. Welsh[2]

UNSW Data Science Hub, School of Mathematics and Statisics, University of New South Wales[1]

Research School of Finance, Actuarial Studies and Statistics, Australian National University[2]

Abstract:

Estimating characteristics of domains (referred to as small areas) within a population from a sample survey of the population is an important problem in survey statistics. In this paper, we consider model-based small area estimation under the nested error regression model. We discuss the construction of mixed model estimators (empirical best linear unbiased predictors, EBLUPs) of small area means and the conditional linear predictors of small area means. Under the asymptotic framework of increasing numbers of small areas and increasing numbers of units in each area, we establish asymptotic linearity results and central limit theorems for these estimators which allow us to establish asymptotic equivalences between estimators, approximate their sampling distributions, obtain simple expressions for and construct simple estimators of their asymptotic mean squared errors, and justify asymptotic prediction intervals. We present model-based simulations that show that in quite small, finite samples, our mean squared error estimator performs as well or better than the widely-used Prasad and Rao (1990) type estimators and is much simpler, so is easier to interpret. We also carry out a design-based simulation using real data on consumer expenditure on fresh milk products to explore the design-based properties of the mixed model estimators. We explain and interpret some surprising simulation results through analysis of the population and further design-based simulations that highlight important differences between the model- and design-based properties of mixed model estimators in small area estimation.

Key words and phrases: Increasing area size asymptotics, indirect estimator, mean squared error estimation, mixed model estimator, model-based prediction, prediction intervals

1. Introduction

Estimates of characteristics such as means and totals for areas, domains or clusters within a population (all referred to as areas) obtained from sample survey data are widely used for resource allocation in social, education and environmental programs, and as the basis for commercial decisions. Direct estimates which use only data specific to an area, can have large standard errors because of relatively small area-specific sample sizes. Small area estimation is concerned with producing more reliable estimates with valid measures of uncertainty for the characteristics of interest; recent reviews include Rao (2005); Jiang and Lahiri (2006b); Rao (2008); Lehtonen and Veijanen (2009); Pfeffermann (2013); Pratesi (2016); Sugasawa and Kubokawa (2020), and Morales et al. (2021).

A popular method for small-area estimation (Fay and Herriot, 1979; Battese et al., 1988) is to introduce a population-level mixed model that includes fixed effects (to describe unit-level and/or area-level effects) and random effects (to describe additional between area variation), fit the model using sample data from multiple areas and then, use the fitted model to construct the desired estimates. For estimating means or totals from unit-level data, a widely used approach (Battese et al., 1988) is to use empirical best linear unbiased predictors (EBLUPs) obtained by minimising the (prediction) mean squared error under the nested error regression or random intercept model and then estimating the unknown quantities by maximum likelihood or restricted maximum likelihood (REML) estimation; see for example Saei and Chambers (2003b,a), Jiang and Lahiri (2006a) and Haslett and Welsh (2019). Some authors (e.g. reference) target the conditional expectations of the small area means given the random effects rather than the means themselves. These two targets are different and have different EBLUPs with potentially different mean squared errors, but are often treated interchangeably in small area estimation. They are both random variables under the model-based framework, so technically they need to be predicted rather than estimated. However, both “prediction” and “estimation” are used in small area estimation so we refer to the EBLUPs as mixed model estimators and distinguish them by their different targets (Tzavidis et al., 2010); they are sometimes called composite and synthetic estimators respectively.

The model-based variability of small area estimators can be reported through estimates of their (prediction) mean squared errors or by prediction intervals. Estimation of (prediction) mean squared errors for mixed model estimators is complicated, even for simple linear mixed models like the nested error regression model.

Under normal linear mixed models (including the nested error regression model), when the number of areas is allowed to increase while the area sizes are held fixed (or bounded), Kackar and Harville (1981) and Prasad and Rao (1990) used Taylor expansions to obtain approximations to the (prediction) mean squared error of the EBLUPs of the conditional expectation of the small area means, and then constructed mean squared error estimators by replacing the unknown quantities in these approximations by estimators; Rao and Molina (2015) later derived a modified approximation for estimating the small area mean. The Prasad-Rao approximation and estimator have been extended to other models and to allow additional estimators of the model parameters by Datta and Lahiri (2000) and Das et al. (2004); see also ZÄĚdÅĆo (2009) and Torabi and Rao (2013). The area-level jackknife (treating the small area means as the characteristics of interest) (Jiang et al., 2002) is an alternative to the analytic approximations. Chatterjee et al. (2008) proposed a parametric bootstrap approach for constructing prediction intervals.

The standard asymptotic framework for model-based small area estimation under the nested error regression model follows Kackar and Harville (1981) and Prasad and Rao (1990) in allowing the number of areas to increase while holding the area sizes fixed (or bounded). In this framework, small area estimators are not consistent and there are no asymptotic distribution results. Consequently, the construction of mean squared error estimates is complicated and prediction intervals based on the estimated mean squared errors cannot be shown to achieve their nominal level even asymptotically. To overcome these difficulties, we need both the number of areas and the sample size in each area to increase. This appears to contradict the “small” in “small area estimation”. However, i) the framework is directly relevant to applications that include a number of large “small areas” and no tiny ones and ii) all asymptotic results are simply a mathematically rigorous way of obtaining approximations to use with (even very small) finite samples. i.e. outside the strict framework. Approximations derived within the increasing small area size framework perform well in other contexts even when some areas have quite small sample size, e.g., for 10 areas with area size 10 (Lyu and Welsh, 2022a,b), and we will show good results with even smaller sample sizes later in this paper. Thus the results can be used successfully in problems that do not appear to fit the framework. Examples of problems that do fit the framework occur in clinical research (clustered trials) when we study records on large groups (areas) of patients (units) with each group treated by a different medical practitioner or at a different hospital, in educational research when we look at records on college students (units) grouped within schools (areas), and in sample surveys when we observe people or households (units) grouped in defined clusters (areas). For example, Arora and Lahiri (1997) gave an example with 43 areas ranging in size from 95 to 633 units, and such examples are common in poverty data (Pratesi, 2016). In the era of big data, examples where small area estimation and our asymptotic framework are directly relevant are increasingly likely to occur. Although direct estimation is then a feasible alternative, small area estimation techniques still offer opportunities to extract additional, more precise information from the data, and thereby enable us to address complex challenges and improve decision-making processes. For example, in public health surveillance, we may want to estimate health-related variables, such as disease incidence and access to healthcare services, to enable health authorities to better identify health disparities and allocate resources more effectively. Similarly, in precision agriculture, we may want to analyze high-resolution data on crops, soil, and weather across smaller zones within farmlands to enable accurate and precise agricultural management practices, ultimately benefiting farmers and the environment.

This study fills current practical and theoretical gaps in small area estimation by deriving the asymptotic distributions of some model-based small area estimators (EBLUPs) of area means and conditional linear predictors. Without assuming normality, we obtain straightforward approximations to the estimators' distributions that include the distribution of the target of interest and a normal distribution. These approximations yield simple, easily estimated expressions for the asymptotic mean squared errors of the estimators and enable the construction of prediction intervals with proven accurate asymptotic coverage. The present work is related to that of Lyu and Welsh (2022a) who also used the asymptotic results for maximum likelihood and restricted maximum likelihood (REML) estimators of the parameters in mixed model obtained by Lyu and Welsh (2022b), but considered estimation of the random effects in the model instead of small area estimation. There are common issues between the two problems (e.g. they are both prediction problems) but the targets are different (unobservable random effects versus potentially observable finite population parameters) so their estimators are different, have different asymptotic distributions and therefore different asymptotic mean squared errors and confidence intervals.

We describe the nested error regression model, discuss the targets of estimation and the mixed model estimators we consider in Section 2. We present our increasing number of areas and increasing area size asymptotic results in Section 3 and use (model-based) simulation to demonstrate the relevance of these results to finite samples in Section 4. We include a design-based simulation using real consumer expenditure on fresh milk products data, and then use additional design-based simulations to explore some unexpected findings in Section 5. We conclude with a brief discussion in Section 6.

2. Small area estimation

Consider a population $U = \cup_{i=1}^{g} U_i$ of N units, partitioned into g exclusive areas U_i, each containing N_i units so that $\sum_{i=1}^{g} N_i = N$. Let y_{ij} be a scalar survey variable of interest and $\mathbf{x}_{ij}$ be a vector of auxiliary variables for the jth unit in the ith area. The problem of interest is to use data from a sample of units in U to make inference about the area means $\bar{y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$. We assume that the values of the auxiliary variables are known for every unit in the sample, the population area means of the auxiliary variables are known, and the values of the survey variable are observed on a sample of units $s = \cup_{i=1}^{g} s_i$, where $s_i \subseteq U_i$ of size $n_i \leq N_i$ is the set of sample units within U_i and $\sum_{i=1}^{g} n_i = n$. We assume that the units are selected using a non-informative sampling method such that the minimum area sample size $n_L > 0$, so that some units from every area are included in the sample. The available data $\mathscr{D}$ consists of $(y_{ij}, \mathbf{x}_{ij})$ for $j \in s_i$, $i = 1, \ldots, g$ and $N_i^{-1} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ for $i = 1, \ldots, g$.

We assume the nested error regression (or random intercept) model for the survey variable at the population level, so

$$y_{ij} = \mu(\mathbf{x}_{ij}) + \alpha_i + e_{ij}, \qquad \text{for } j = 1, \ldots, N_i,\ i = 1, \ldots, g, \tag{2.1}$$

where $\mu(\mathbf{x}_{ij})$ is the regression function (the conditional mean of the response) given $\mathbf{x}_{ij}$, α_i is a random effect representing a random intercept or area effect and e_{ij} is a random error. We assume that the $\{\alpha_i\}$ and $\{e_{ij}\}$ are all mutually independent with mean zero and variances (called variance components) σ_α^2 and σ_e^2, respectively, and write $\boldsymbol{\theta} = [\sigma_\alpha^2, \sigma_e^2]^T$. These random variables do not have to be normally distributed so the responses y_i are not necessarily normally distributed. Lyu and Welsh (2022b) showed that, in asymptotic theory with increasing area size, we need to distinguish within area variables (unit level variables that vary within areas so need subscripts i and j) which we place in the p_w-vector $\mathbf{x}_{ij}^{(w)}$ and between area variables (area level variables that are constant within areas so only need subscript i) which we place in the p_b-vector $\mathbf{x}_i^{(b)}$. Distinguishing the two kinds of variables is important because they contain different amounts of information and hence estimators of their coefficients have different rates of convergence. We then write the regression function as

$$\mu(\mathbf{x}_{ij}) = \beta_0 + \mathbf{x}_i^{(b)T}\boldsymbol{\beta}_1 + \mathbf{x}_{ij}^{(w)T}\boldsymbol{\beta}_2 = \mathbf{u}_i^T\boldsymbol{\xi} + \mathbf{x}_{ij}^{(w)T}\boldsymbol{\beta}_2, \tag{2.2}$$

where β_0 is the unknown intercept, $\boldsymbol{\beta}_1$ is the unknown between area slope and $\boldsymbol{\beta}_2$ is the unknown within area slope. It is convenient to group the intercept and the between area slope terms in the $(p_b+1)-$vectors $\mathbf{u}_i = [1,\ \mathbf{x}_i^{(b)T}]^T$ and $\boldsymbol{\xi} = [\beta_0, \boldsymbol{\beta}_1^T]^T$. It is also sometimes useful to write the regression function in terms of $\mathbf{z}_{ij} = [1,\ \mathbf{x}_i^{(b)T},\ \mathbf{x}_{ij}^{(w)T}]^T$ $= [\mathbf{u}_i^T,\ \mathbf{x}_{ij}^{(w)T}]^T$ and $\boldsymbol{\beta} = [\beta_0, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T]^T = [\boldsymbol{\xi}^T, \boldsymbol{\beta}_2^T]^T$. Finally, grouping the between and within area parameters, the full set of model parameters is $\boldsymbol{\omega} = [\beta_0, \boldsymbol{\beta}_1^T, \sigma_\alpha^2, \boldsymbol{\beta}_2^T, \sigma_e^2]^T$.

The regression function (2.2) includes the three cases with either or both between and within area variables. It also gives us the option of replacing $\mathbf{x}_{ij}^{(w)}$ by the population-area-mean-centered within area variables $\mathbf{x}_{ij}^{(w)} - \bar{\mathbf{x}}_i^{(w)}$, where $\bar{\mathbf{x}}_i^{(w)} = N_i^{-1}\sum_{k=1}^{N_i}\mathbf{x}_{ik}^{(w)}$, for $j = 1,\ldots,N_i$, $i = 1,\ldots,g$, and including the area means $\bar{\mathbf{x}}_i^{(w)}$ as contextual effects with the between area variables from $\mathbf{x}_{ij}$ in $\mathbf{x}_i^{(b)}$. This centering makes the between and within area variables orthogonal and allows a simple interpretation of the parameters; this is the reason we center about the population area means rather than the sample area means. See Yoon and Welsh (2020) and the references therein for more discussion of the benefits of centering the within area variables.

The two common targets when we are interested in the small area means are the actual small area means $\bar{y}_i = \mathbf{u}_i^T \boldsymbol{\xi} + \bar{\mathbf{x}}_i^{(w)^T} \boldsymbol{\beta}_2 + \alpha_i + \bar{e}_i$ and the conditional linear predictors of the small area means

$$\eta_i = \mathrm{E}(\bar{y}_i | \mathbf{u}_i, \mathbf{x}_{i1}^{w}, \ldots, \mathbf{x}_{iN_i}^{(w)}, \alpha_i) = \mathbf{u}_i^T \boldsymbol{\xi} + \bar{\mathbf{x}}_i^{(w)T} \boldsymbol{\beta}_2 + \alpha_i, \quad i = 1, \ldots, g.$$

The targets have the same expectations $\mathrm{E}(\bar{y}_i) = \mathrm{E}(\eta_i)$ but different variances $\mathrm{Var}(\bar{y}_i) = \sigma_\alpha^2 + N_i^{-1}\sigma_e^2$ and $\mathrm{Var}(\eta_i) = \sigma_\alpha^2$. This means that unbiased predictors of one will also be unbiased predictors of the other, but the (prediction) mean squared errors of unbiased predictors will differ with the target. We prefer to predict $\bar{y}_i$ rather than η_i because $\bar{y}_i$ is a simple finite population parameter whereas η_i is tied to the specific population model we are using and hence is more difficult to interpret than $\bar{y}_i$. Also, when U_i is completely enumerated, $\bar{y}_i$, unlike η_i, can be evaluated without any prediction error. In the standard asymptotic framework with fixed area size, the difference between the targets is fixed; under our increasing area size framework, the two targets are asymptotically the same up to order $N_i^{-1/2}$, because

$$\bar{y}_i - \eta_i = \bar{e}_i = O_p(N_i^{-1/2}), \qquad \text{as } N_i \to \infty,$$

where O_p denotes stochastic boundedness in probability. This means that the predictors are quite similar in large areas and may explain why the distinction between the two targets has been largely ignored in practice.

The prediction mean squared error for predicting a target random variable is minimised by the conditional expectation of the target given the observed data $\mathscr{D}$. Typically, the distribution of the target given $\mathscr{D}$ is derived from a model with unknown parameters (such as (2.1) and (2.2)) so the conditional expectation depends on these unknown parameters and a feasible predictor requires replacing the unknown parameters by estimators; in our case, we use the maximum likelihood or REML estimators of the parameters in the model (2.1) and (2.2).

To simplify notation, we write $j \not\in s_i$ to mean $j \in U_i \backslash s_i$, $k_i = (N_i - n_i)/N_i$, and use subscripts (s) and (r) to denote quantities related to sampled and non-sampled units. Specifically, let $\bar{y}_{i(s)} = n_i^{-1}\sum_{j\in s_i} y_{ij}$, $\bar{y}_{i(r)} = (N_i - n_i)^{-1}\sum_{j\not\in s_i} y_{ij}$, $\bar{\mathbf{x}}_{i(s)}^{(w)} = n_i^{-1}\sum_{j\in s_i}\mathbf{x}_{ij}^{(w)}$, $\bar{\mathbf{x}}_{i(r)}^{(w)} = (N_i - n_i)^{-1}\sum_{j\not\in s_i}\mathbf{x}_{ij}^{(w)}$, $\bar{e}_{i(s)} = n_i^{-1}\sum_{j\in s_i} e_{ij}$ and $\bar{e}_{i(r)} = (N_i - n_i)^{-1}\sum_{j\not\in s_i} e_{ij}$. Then, under the model (2.1) and (2.2), we can write the actual small area means $\bar{y}_i$ as

$$\bar{y}_i = (1-k_i)\bar{y}_{i(s)} + k_i\bar{y}_{i(r)} = (1-k_i)\bar{y}_{i(s)} + k_i(\mathbf{u}_i^T\boldsymbol{\xi} + \bar{\mathbf{x}}_{i(r)}^{(w)T}\boldsymbol{\beta}_2 + \alpha_i + \bar{e}_{i(r)}). \qquad (2.3)$$

Taking the conditional expectation given the data $\mathscr{D}$ of (2.3) and substituting the maximum likelihood or restricted maximum likelihood (REML) estimators $\hat{\boldsymbol{\xi}}$ and $\hat{\boldsymbol{\beta}}_2$ for $\boldsymbol{\xi}$ and $\boldsymbol{\beta}_2$, respectively, and predicting $\mathrm{E}(\alpha_i|\mathscr{D})$ by the empirical best linear unbiased predictor (EBLUP)

$$\hat{\alpha}_i = \hat{\gamma}_i\{\bar{y}_{i(s)} - \mathbf{u}_i^T\hat{\boldsymbol{\xi}} - \bar{\mathbf{x}}_{i(s)}^{(w)T}\hat{\boldsymbol{\beta}}_2\}, \qquad \text{with } \hat{\gamma}_i = n_i\hat{\sigma}_\alpha^2/(\hat{\sigma}_e^2 + n_i\hat{\sigma}_\alpha^2), \qquad (2.4)$$

where $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_e^2$ are maximum likelihood or REML estimators of σ_α^2 and σ_e^2, respectively, we obtain the predictor

$$\hat{M}_i^{sam} = (1-k_i)\bar{y}_{i(s)} + k_i\{\mathbf{u}_i^T\hat{\boldsymbol{\xi}} + \bar{\mathbf{x}}_{i(r)}^{(w)T}\hat{\boldsymbol{\beta}}_2 + \hat{\alpha}_i\}, \qquad (2.5)$$

which is the mixed model estimator of $\bar{y}_i$ (Tzavidis et al., 2010); the superscript ’sam’ shows that the target is the ‘small area mean’. It can also be called an EBLUP or a composite estimator (Costa et al., 2003) as it combines the sample mean $\bar{y}_{i(s)}$ with a synthetic component $\mathbf{u}_i^T\hat{\boldsymbol{\xi}}+\bar{\mathbf{x}}_{i(r)}^{(w)T}\hat{\boldsymbol{\beta}}_2+\hat{\alpha}_i$ (Rao, 2008). The mixed model estimator of η_i obtained similarly is the EBLUP

$$\hat{M}_i^{\mathrm{clp}} = \mathbf{u}_i^T\hat{\boldsymbol{\xi}}+\bar{\mathbf{x}}_i^{(w)T}\hat{\boldsymbol{\beta}}_2+\hat{\alpha}_i = (1-k_i)\{\mathbf{u}_i^T\hat{\boldsymbol{\xi}}+\bar{\mathbf{x}}_{i(s)}^{(w)T}\hat{\boldsymbol{\beta}}_2+\hat{\alpha}_i\}+k_i\{\mathbf{u}_i^T\hat{\boldsymbol{\xi}}+\bar{\mathbf{x}}_{i(r)}^{(w)T}\hat{\boldsymbol{\beta}}_2+\hat{\alpha}_i\}, \tag{2.6}$$

which is a fully synthetic or indirect estimator (Prasad and Rao, 1990; Lahiri and Rao, 1995; Jiang et al., 2011); the superscript ‘clp’ shows that the target is the ‘conditional linear predictor’.

The main difference between the estimators (2.5) of $\bar{y}_i$ and (2.6) of η_i is that the former uses the observed $\bar{y}_{i(s)}$ whereas the latter uses a model-based prediction for this quantity. The difference between the estimators of the two targets can be expressed in terms of the EBLUP $\hat{\alpha}_i$ for the random effect (2.4) as

$$\hat{M}_i^{\mathrm{sam}}-\hat{M}_i^{\mathrm{clp}} = (1-k_i)\{\bar{y}_{i(s)}-\mathbf{u}_i^T\hat{\boldsymbol{\xi}}-\bar{\mathbf{x}}_{i(s)}^{(w)T}\hat{\boldsymbol{\beta}}_2-\hat{\alpha}_i\} = \frac{n_i}{N_i}\left\{\frac{\hat{\alpha}_i}{\hat{\gamma}_i}-\hat{\alpha}_i\right\} = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_\alpha^2}\frac{\hat{\alpha}_i}{N_i}. \tag{2.7}$$

This difference is often quite small, but it can be large for areas with extreme EBLUPs $\hat{\alpha}_i$, particularly if N_i is small and the estimated within area variance is much larger than the estimated between area variance so $\hat{\sigma}_e^2 > \hat{\sigma}_\alpha^2$. Asymptotically, the difference is $O_p(N_i^{-1})$, so the estimators are asymptotically equivalent up to this order and hence asymptotically closer than their respective targets; see Section 3 for details. Again, these properties only hold when the area sizes are increasing and do not hold for fixed area-size asymptotics.

We emphasise that the predictor depends on the target random variable, the data and the model. If, instead of $\mathscr{D}$, we only observe the sample data $(y_{ij}, \mathbf{x}_{ij}^T)^T$, $j = 1, \dots n_i$, $i = 1, \dots, g$, the population mean $\bar{\mathbf{x}}_i^{(w)}$ and hence the non-sample mean of the within area variables $\bar{\mathbf{x}}_{i(r)}^{(w)}$ is unknown and also needs to be predicted. If we estimate $\bar{\mathbf{x}}_{i(r)}^{(w)}$ by the simple nonparametric estimator $\bar{\mathbf{x}}_{i(s)}^{(w)}$, we obtain the mixed model estimator/predictor of the small area mean $\bar{y}_i$ given by $\hat{M}_i^{*sam} = (1-k_i)\bar{y}_{i(s)} + k_i\{\mathbf{u}_i^T\hat{\boldsymbol{\xi}} + \bar{\mathbf{x}}_{i(s)}^{(w)T}\hat{\boldsymbol{\beta}}_2 + \hat{\alpha}_i\}$.

Theorem 1 in Section 3 below establishes that, as $g, n_L \to \infty$, the mixed model estimators (2.5) and (2.6) are asymptotically equivalent predictors of the small area means $\bar{y}_i$; Theorem 2 establishes that they are asymptotically unbiased predictors and their asymptotic prediction mean squared errors are $\text{MSE}_{\text{LW},i} = n_i^{-1}k_i\sigma_e^2$. We can estimate $\text{MSE}_{\text{LW},i}$ by substituting the consistent maximum likelihood or REML estimator $\hat{\sigma}_e^2$ for σ_e^2 to obtain

$$\widehat{\text{MSE}}_{\text{LW},i} = n_i^{-1}k_i\hat{\sigma}_e^2. \tag{2.8}$$

We can then construct simple, asymptotic $100(1-\varepsilon)\%$ prediction intervals for $\bar{y}_i$ which we denote sam-LW and clp-LW, respectively, as

$$[\hat{M}_i^{\text{sam}} - \Phi^{-1}(1-\varepsilon/2)\widehat{\text{MSE}}_{\text{LW},i}^{1/2}, \hat{M}_i^{\text{sam}} + \Phi^{-1}(1-\varepsilon/2)\widehat{\text{MSE}}_{\text{LW},i}^{1/2}], \tag{2.9}$$

$$[\hat{M}_i^{\text{clp}} - \Phi^{-1}(1-\varepsilon/2)\widehat{\text{MSE}}_{\text{LW},i}^{1/2}, \hat{M}_i^{\text{clp}} + \Phi^{-1}(1-\varepsilon/2)\widehat{\text{MSE}}_{\text{LW},i}^{1/2}], \tag{2.10}$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function. The asymptotic coverage of the intervals (2.9) and (2.10) is guaranteed by Lyu and Welsh (2022b), Slutsky's Theorem and Theorem 2, which do not require the assumption of normality in the model.

3. Increasing area-size asymptotic results

We assume throughout that the true model describing the actual data generating mechanism is given by (2.1) and (2.2) with true parameter $\dot{\boldsymbol{\omega}} = [\dot{\beta}_0, \dot{\boldsymbol{\beta}}_1^T, \dot{\sigma}_\alpha^2, \dot{\boldsymbol{\beta}}_2^T, \dot{\sigma}_e^2]^T$ and take all expectations under the true model. Let $\hat{\boldsymbol{\omega}}$ denote the normal maximum likelihood estimator (MLE) of $\dot{\boldsymbol{\omega}}$ obtained by maximizing the normal likelihood based on a sample s. Similarly, let $\hat{\boldsymbol{\theta}}_R$ be the normal REML estimator of $\dot{\boldsymbol{\theta}} = [\dot{\sigma}_\alpha^2, \dot{\sigma}_e^2]^T$ obtained by maximizing the normal REML criterion function and let $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = [\hat{\beta}_0(\boldsymbol{\theta}), \hat{\boldsymbol{\beta}}_1(\boldsymbol{\theta})^T, \hat{\boldsymbol{\beta}}_2(\boldsymbol{\theta})^T]^T$ be the profile likelihood estimator of $\dot{\boldsymbol{\beta}} = [\dot{\beta}_0, \dot{\boldsymbol{\beta}}_1^T, \dot{\boldsymbol{\beta}}_2^T]^T$ obtained by maximizing the normal likelihood with $\boldsymbol{\theta}$ held fixed. We call $\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}_R)$ the normal REML estimator of $\dot{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\omega}}_R = (\hat{\beta}_{R0}, \hat{\boldsymbol{\beta}}_{R1}^T, \hat{\sigma}_{R\alpha}^2, \hat{\boldsymbol{\beta}}_{R2}^T, \hat{\sigma}_{Re}^2)^T$ the normal REML estimator of $\dot{\boldsymbol{\omega}}$.

Following Lyu and Welsh (2022b), we impose the following conditions:

Condition A

1. The model (2.1) and (2.2) holds with true parameters $\dot{\boldsymbol{\omega}}$ inside the parameter space Ω.

2. The number of areas $g \to \infty$ and minimum area sample size $n_L \to \infty$.

3. The random variables $\{\alpha_i\}$ and $\{e_{ij}\}$ are independent and identically distributed and there is a $\delta > 0$ such that $\mathrm{E}|\alpha_i|^{4+\delta} < \infty$ and $\mathrm{E}|e_{ij}|^{4+\delta} < \infty$ for all $j = 1, \ldots, N_i$, $i = 1, \ldots, g$.

4. The limits $\mathbf{b}_1 = \lim_{g\to\infty} g^{-1} \sum_{i=1}^{g} \mathbf{x}_i^{(b)}$, $\mathbf{B}_2 = \lim_{g\to\infty} g^{-1} \sum_{i=1}^{g} \mathbf{x}_i^{(b)} \mathbf{x}_i^{(b)^T}$ and $\mathbf{B}_3 = \lim_{g\to\infty} \lim_{n_L\to\infty} n^{-1} \sum_{i=1}^{g} \sum_{j\in s_i} (\mathbf{x}_{ij}^{(w)} - \bar{\mathbf{x}}_i^{(w)})(\mathbf{x}_{ij}^{(w)} - \bar{\mathbf{x}}_i^{(w)})^T$ exist, and the matrices $\mathbf{B}_2$ and $\mathbf{B}_3$ are positive definite. Further, $\lim_{g\to\infty} g^{-1} \sum_{i=1}^{g} |\bar{\mathbf{x}}_i^{(w)}|^2 < \infty$, and there exists a $\delta > 0$ such that $\lim_{g\to\infty} g^{-1} \sum_{i=1}^{g} |\mathbf{x}_i^{(b)}|^{2+\delta} < \infty$ and $\lim_{g\to\infty} \lim_{n_L\to\infty} n^{-1} \sum_{i=1}^{g} \sum_{j\in s_i} |\mathbf{x}_{ij}^{(w)} - \bar{\mathbf{x}}_i^{(w)}|^{2+\delta} < \infty$.

As noted in Lyu and Welsh (2022b), these are mild conditions. Conditions A3 and A4 ensure that limits needed to ensure the existence of the asymptotic variance of the estimating function exist, and that we can establish a Lyapounov condition and hence a central limit theorem for the estimating function. Condition A4 ensures the matrix

$$\mathbf{B} = \text{block diag}[\mathbf{B}_u/\dot{\sigma}_\alpha^2, 1/(2\dot{\sigma}_\alpha^4), \mathbf{B}_3/\dot{\sigma}_e^2, 1/(2\dot{\sigma}_e^4)], \qquad \text{with } \mathbf{B}_u = \begin{bmatrix} 1 & \mathbf{b}_1^T \\ \mathbf{b}_1 & \mathbf{B}_2 \end{bmatrix},$$

is positive definite. For later, note that $\mathbf{B}_u = \lim_{g\to\infty} g^{-1} \sum_{i=1}^{g} \mathbf{u}_i \mathbf{u}_i^T$.

We use the central limit theorem of Lyu and Welsh (2022b) for the maximum likelihood and REML estimators to derive the asymptotic distribution of the mixed model estimators (2.5) and (2.6). These results are achieved by approximating the estimators directly and taking the (prediction) mean squared error of the approximation, rather than directly approximating the (prediction) mean squared error. Consequently, we begin by establishing asymptotic linearity results for the estimators. Recall that $k_i = (N_i - n_i)/N_i$.

Theorem 1. *Suppose Condition A holds. Then, as $g, n_L \to \infty$, we have*

$$\hat{M}_i^{sam} - \bar{y}_i = k_i\{\bar{e}_{i(s)} - \bar{e}_{i(r)}\} + k_i O_p(n_L^{-1} + g^{-1/2} n_L^{-1/2}),$$

$$\hat{M}_i^{clp} - \bar{y}_i = k_i\{\bar{e}_{i(s)} - \bar{e}_{i(r)}\} + O_p(n_L^{-1} + g^{-1/2} n_L^{-1/2}) \quad \textit{and}$$

$$\hat{M}_i^{clp} - \dot{\eta}_i = \bar{e}_{i(s)} + O_p(n_L^{-1} + g^{-1/2} n_L^{-1/2}).$$

Proof. From (2.5) and (2.3), we can write

$$\hat{M}_i^{\text{sam}} - \bar{y}_i = k_i \left\{ \mathbf{u}_i^T(\hat{\boldsymbol{\xi}} - \dot{\boldsymbol{\xi}}) + \bar{\mathbf{x}}_{i(r)}^{(w)T}(\hat{\boldsymbol{\beta}}_2 - \dot{\boldsymbol{\beta}}_2) + \hat{\alpha}_i - \alpha_i + \bar{e}_{i(r)} \right\}.$$

Using the approximation

$$\hat{\alpha}_i = \alpha_i + \bar{e}_{i(s)} - \mathbf{u}_i^T(\hat{\boldsymbol{\xi}} - \dot{\boldsymbol{\xi}}) + O_p(n_L^{-1} + g^{-1/2} n_L^{-1/2}) \tag{3.11}$$

obtained by Lyu and Welsh (2022a) in the proof of their Theorem 2 (see also their Supplementary Material page 14), and the result from the central limit theorem of Lyu and Welsh (2022b) that $\hat{\boldsymbol{\beta}}_2 - \dot{\boldsymbol{\beta}}_2 = O_p(n^{-1/2}) = O_p(g^{-1/2} n_L^{-1/2})$, as $g, n_L \to \infty$, we obtain the approximation

$$\begin{aligned} \hat{M}_i^{\text{sam}} - \bar{y}_i &= k_i \left\{ \bar{e}_{i(s)} - \bar{e}_{i(r)} + \bar{\mathbf{x}}_{i(r)}^{(w)T}(\hat{\boldsymbol{\beta}}_2 - \dot{\boldsymbol{\beta}}_2) + O_p(n_L^{-1} + g^{-1/2} n_L^{-1/2}) \right\} \\ &= k_i\{\bar{e}_{i(s)} - \bar{e}_{i(r)}\} + k_i O_p(n_L^{-1} + g^{-1/2} n_L^{-1/2}). \end{aligned}$$

Similarly, from (2.6) and (2.3), we have

$$\hat{M}_i^{\text{clp}} - \bar{y}_i = \mathbf{u}_i^T(\hat{\boldsymbol{\xi}} - \dot{\boldsymbol{\xi}}) + \bar{\mathbf{x}}_i^{(w)T}(\hat{\boldsymbol{\beta}}_2 - \dot{\boldsymbol{\beta}}_2) + (\hat{\alpha}_i - \alpha_i) - \bar{e}_i.$$

Substituting the approximation (3.11) for $\hat{\alpha}_i$, we then obtain

$$\begin{aligned} \hat{M}_i^{\text{clp}} - \bar{y}_i &= \bar{e}_{i(s)} - \bar{e}_i + \bar{\mathbf{x}}_{i(r)}^{(w)T}(\hat{\boldsymbol{\beta}}_2 - \dot{\boldsymbol{\beta}}_2) + O_p(n_L^{-1} + g^{-1/2}n_L^{-1/2}) \\ &= \bar{e}_{i(s)} - \bar{e}_i + O_p(n_L^{-1} + g^{-1/2}n_L^{-1/2}) \qquad (3.12) \\ &= k_i\{\bar{e}_{i(s)} - \bar{e}_{i(r)}\} + O_p(n_L^{-1} + g^{-1/2}n_L^{-1/2}), \end{aligned}$$

because $k_i\bar{e}_{i(s)} - k_i\bar{e}_{i(r)} = k_i\bar{e}_{i(s)} - \{\bar{e}_i - (1-k_i)\bar{e}_{i(s)}\} = \bar{e}_{i(s)} - \bar{e}_i$.

The final result follows from the fact that $\bar{y}_i = \dot{\eta}_i + \bar{e}_i$, so we have

$$\hat{M}_i^{\text{clp}} - \dot{\eta}_i = \hat{M}_i^{\text{clp}} - (\bar{y}_i - \bar{e}_i) = \bar{e}_{i(s)} + O_p(n_L^{-1} + g^{-1/2}n_L^{-1/2}),$$

using (3.12). □

We can also consider using the estimator of the small area mean to estimate $\dot{\eta}_i$. We obtain the same leading term as for the estimator of the conditional linear predictor because $k_i\{\bar{e}_{i(s)} - \bar{e}_{i(r)}\} + \bar{e}_i = \bar{e}_{i(s)} - (n_i/N_i)\bar{e}_{i(s)} - k_i\bar{e}_{i(r)} + \bar{e}_i = \bar{e}_{i(s)}$. However, it would be unusual in practice to use the estimator of the small area mean for this purpose, so we only state the formal result for the estimator of the conditional linear predictor. The estimators of the small area mean and the conditional linear predictor have the same leading terms (so are asymptotically equivalent to first order). However the remainders for the two estimators are different, showing that there can be higher order differences between them. In particular, If we sample the entire ith area (so $n_i = N_i$ and $k_i = 0$), we have $\hat{M}_i^{\text{sam}} - \bar{y}_i = 0$ but $\hat{M}_i^{\text{clp}} - \bar{y}_i = O_p(n_L^{-1} + g^{-1/2}n_L^{-1/2})$, so $\hat{M}_i^{\text{clp}} - \bar{y}_i$ is only asymptotically zero.

The asymptotic distribution of the estimators is given by the following theorem.

Theorem 2. *Suppose Condition A holds, $n_i^{1/2}/n_L \to 0$ and $n_i/N_i \to f_i$ for $0 \le f_i < 1$. Then as $g, n_L \to \infty$, we have*

$$n_i^{1/2}(\hat{M}_i^{sam} - \bar{y}_i) \xrightarrow{D} N(0, (1-f_i)\dot{\sigma}_e^2), \quad n_i^{1/2}(\hat{M}_i^{clp} - \bar{y}_i) \xrightarrow{D} N(0, (1-f_i)\dot{\sigma}_e^2),$$

$$\textit{and} \qquad n_i^{1/2}(\hat{M}_i^{clp} - \dot{\eta}_i) \xrightarrow{D} N(0, \dot{\sigma}_e^2).$$

Proof. Write $\bar{e}_{i(s)} - \bar{e}_{i(r)} = \sum_{j=1}^{N_i} w_{ij}e_{ij}$, where $w_{ij} = n_i^{-1}I_{ij} - (N_i - n_i)^{-1}(1 - I_{ij})$ with $I_{ij} = 1$ if unit j in area i is selected in the sample and 0 otherwise. Using the fact that the e_{ij} are independent, we can show that $\text{Var}(\sum_{j=1}^{N_i} w_{ij}e_{ij}) = (n_ik_i)^{-1}\dot{\sigma}_e^2$ and, for $\delta > 0$ such that $\text{E}|e_{ij}|^{2+\delta} < \infty$, $\sum_{j=1}^{N_i}\text{E}|w_{ij}e_{ij}|^{2+\delta} = O(n_i^{-1-\delta})$. It follows that the Lyapounov condition holds, and hence from the central limit theorem that

$$(n_ik_i)^{1/2}\sum_{j=1}^{N_i} w_{ij}e_{ij} \xrightarrow{D} N(0, \dot{\sigma}_e^2).$$

Since $f_i < 1$, the theorem follows from Theorem 1. The final statement follows because $n_i^{1/2}\bar{e}_{i(s)} \xrightarrow{D} N(0, \dot{\sigma}_e^2)$. □

Theorem 2 allows us to reach several interesting conclusions.

i) The asymptotic distribution of both estimators is the distribution of the target characteristic of interest $F_{\bar{y}_i}$ or $F_{\dot{\eta}_i}$; Theorem 2 suggests that a better approximation is the distribution of $k_i(\bar{e}_{i(s)} - \bar{e}_{i(r)})$ plus $\bar{y}_i$ when the target is

$\bar{y}_i$, or the distribution of $\bar{e}_{i(s)}$ plus $\dot{\eta}_i$ when the target is $\dot{\eta}_i$. The first pair of random variables are uncorrelated while the second are independent (so the distribution is the convolution of the $N(0, n_i^{-1}\dot{\sigma}_e^2)$ distribution with $F_{\dot{\eta}_i}$). These distributions are not the same in general, but when the random effects and error in the model are all normally distributed, these approximations are the same, being $N(\mathbf{u}_i^T\dot{\boldsymbol{\xi}} + \bar{\mathbf{x}}_i^{(w)T}\dot{\boldsymbol{\beta}}_2, \dot{\sigma}_\alpha^2 + N_i^{-1}\dot{\sigma}_e^2 + (1-f_i)n_i^{-1}\dot{\sigma}_e^2) = N(\mathbf{u}_i^T\dot{\boldsymbol{\xi}} + \bar{\mathbf{x}}_i^{(w)T}\dot{\boldsymbol{\beta}}_2, \dot{\sigma}_\alpha^2 + n_i^{-1}\dot{\sigma}_e^2)$ and $N(\mathbf{u}_i^T\dot{\boldsymbol{\xi}} + \bar{\mathbf{x}}_i^{(w)T}\dot{\boldsymbol{\beta}}_2, \dot{\sigma}_\alpha^2 + n_i^{-1}\dot{\sigma}_e^2)$, respectively.

ii) The asymptotic mean squared error of $\hat{M}_i^{\text{clp}}$ for estimating $\dot{\eta}_i$ is $n_i^{-1}\dot{\sigma}_e^2$ which is greater than or equal to $n_i^{-1}(1-f_i)\dot{\sigma}_e^2$, the asymptotic mean squared error of $\hat{M}_i^{\text{clp}}$ (or $\hat{M}_i^{\text{sam}}$) for estimating $\bar{y}_i$. Thus, using the asymptotic mean squared error of $\hat{M}_i^{\text{clp}}$ for estimating $\dot{\eta}_i$ when we are estimating $\bar{y}_i$ is conservative.

iii) We can also use the central limit theorem of Lyu and Welsh (2022b) to describe the rate at which we can estimate the asymptotic mean squared errors. For example, we have $\widehat{\text{MSE}}_{\text{LW},i} - \text{MSE}_{\text{LW},i} = n_i^{-1}k_i(\hat{\sigma}_e^2 - \dot{\sigma}_e^2)$, so it follows from Theorem 1 that

$$k_i^{-1/2}n_i^{1/2}n^{1/2}\left(\widehat{\text{MSE}}_{\text{LW},i} - \text{MSE}_{\text{LW},i}\right) = n^{1/2}(\hat{\sigma}_e^2 - \dot{\sigma}_e^2) \xrightarrow{D} N(0, \mathrm{E}\, e_{11}^4 - \dot{\sigma}_e^4).$$

iv) Theorem 2 establishes that the prediction intervals (2.9) and (2.10) have the correct asymptotic level.

v) The result for estimating $\dot{\eta}_i$ also holds when $f_i = 1$. In this case, for estimating $\bar{y}_i$, provided $N_i - n_i \to \infty$, we have $(N_i - n_i)^{1/2}(\hat{M}_i^{\text{sam}} - \bar{y}_i) \xrightarrow{P} 0$ and $(N_i -$

$n_i)^{1/2}(\hat{M}_i^{\text{clp}} - \bar{y}_i) \xrightarrow{P} 0.$

4. Simulation study

We carried out a model-based simulation study to evaluate the performance of the prediction intervals (2.9) and (2.10). We generated population data with $g \in \{15, 30, 50\}$ small areas and N_i units in each small area by making area 1 the smallest area ($N_1 = N_L = 40$) and then setting the remaining N_i equal to the integer parts of $g-1$ independent uniform $[40, 400]$ random variables. The N_i were generated once for each simulation setting so that each setting involved populations of fixed N_i and hence fixed size $N = \sum_{i=1}^{g} N_i$. For each setting, we generated N population values of an auxiliary variable x_{ij} with an area structure by setting $x_{ij} = 3 + 2u_i + 4v_{ij}$, where u_i and v_{ij} are independent standard normal random variables. We centered the x_{ij} about their small area means $\bar{x}_i$ to obtain the within small area variable $x_{ij} - \bar{x}_i$ and also included $\bar{x}_i$ as a between small area variable. The N population values for y were generated from the model

$$y_{ij} = \beta_0 + \beta_1 \bar{x}_i + \beta_2(x_{ij} - \bar{x}_i) + \alpha_i + e_{ij}, \quad j = 1, \ldots, N_i, \, i = 1, \ldots, g, \qquad (4.13)$$

where $\{\alpha_i\}$ were generated independently from F_α with $\mathrm{E}(\alpha_i) = 0$ and $\mathrm{Var}(\alpha_i) = \sigma_\alpha^2$, and independently, $\{e_{ij}\}$ were generated independently from F_e with $\mathrm{E}(e_{ij}) = 0$ and $\mathrm{Var}(e_{ij}) = \sigma_e^2$. We set the true parameters $\dot{\boldsymbol{\beta}} = [5, 7, 3]^T$, $\dot{\sigma}_\alpha^2 \in \{4, 64\}$ and $\dot{\sigma}_e^2 \in \{25, 100\}$, and the distributions $F_\alpha = N(0, \dot{\sigma}_\alpha^2)$ or $F_\alpha = 0.3N(0.5, 1) + 0.7N\big(\dot{\mu}, \{\dot{\sigma}_\alpha^2 -$

$0.375 - 0.7\dot{\mu}^2\}/0.7\big)$, and $F_e = N(0, \dot{\sigma}_e^2)$ or $F_e = 0.3N(0.5, 1) + 0.7N(\dot{\mu}, \{\dot{\sigma}_e^2 - 0.375 - 0.7\dot{\mu}^2\}/0.7)$ with $\dot{\mu} = -0.3 \times 0.5/0.7$. The 3 values of g and 2 for each of $\dot{\sigma}_\alpha^2$, $\dot{\sigma}_e^2$, F_α and F_e produced 48 different simulation settings.

For each of the 48 simulation settings, we generated 1000 populations and then selected one sample via simple random sampling without replacement from each population. We randomly set three sample sizes $4 \le n_i \le 8$ for three randomly selected areas with population sizes $50 \le N_i \le 200$. For the remaining samples, if $N_i \le 50$, we randomly set n_i between 10 and 20. However, if $50 \le N_i \le 100$, we set $n_i = \lfloor 0.5 * N_i \rfloor$, and if $N_i > 100$, we set $n_i = \lfloor 0.3 * N_i \rfloor$. Here, $\lfloor \quad \rfloor$ denotes the integer part function. We select the units in each area independently through simple random sampling without replacement.

For each sample, we fitted the model (4.13) using REML in `lmer` and computed the 95% prediction intervals based on these estimates described in (2.9) and (2.10). In the results, these intervals are denoted sam-LW and clp-LW to emphasise that the intervals are based on the sam or clp predictors, (2.5) or (2.6), respectively, and the proposed estimator of the LW mean squared error (2.8). For comparison, we computed the Prasad-Rao interval clp-PR based on the clp predictor and the Prasad-Rao estimator PR of the root mean squared error, and the Rao-Molina interval sam-RM based on the sam predictor and the Rao-Molina extension RM of the Prasad-Rao estimator of the root mean squared error for this predictor; see Supplementary Material for details. We also computed some Chatterjee et al. (2008)

Table 1: Simulated coverage and length of prediction intervals when α_i and e_{ij} have normal distributions with variances $\dot{\sigma}_\alpha^2 = 4$ and $\dot{\sigma}_e^2 = 25$, respectively.

Method			Direct		sam-LW		sam-RM		clp-LW		clp-PR	
Area	N_i	n_i	Cvge	Alen	Cvge	Alen	Cvge	Alen	Cvge	Alen	Cvge	Alen
1	74	4	0.994	5.942	0.998	2.431	0.918	1.622	0.998	2.431	0.911	1.604
2	176	5	0.997	5.380	0.993	2.204	0.911	1.535	0.992	2.204	0.910	1.531
3	196	6	0.998	4.927	0.992	2.009	0.935	1.467	0.992	2.009	0.931	1.468
4	44	15	1.000	2.674	0.955	1.048	0.932	0.956	0.949	1.048	0.955	1.112
5	40	20	1.000	2.016	0.961	0.790	0.947	0.750	0.944	0.790	0.978	0.999
6	57	28	1.000	1.732	0.956	0.674	0.943	0.649	0.934	0.674	0.977	0.872
7	64	32	1.000	1.606	0.970	0.625	0.966	0.605	0.960	0.625	0.988	0.824
8	173	52	1.000	1.503	0.949	0.580	0.941	0.564	0.944	0.580	0.974	0.666
9	191	57	1.000	1.435	0.956	0.555	0.953	0.541	0.957	0.555	0.980	0.638
10	225	68	1.000	1.313	0.953	0.506	0.950	0.496	0.944	0.506	0.978	0.588
11	232	70	1.000	1.298	0.951	0.499	0.949	0.490	0.948	0.499	0.980	0.580
12	240	72	1.000	1.273	0.950	0.493	0.943	0.484	0.949	0.493	0.975	0.573
13	252	76	1.000	1.244	0.940	0.480	0.938	0.471	0.937	0.480	0.975	0.559
14	253	76	1.000	1.241	0.944	0.479	0.941	0.471	0.943	0.479	0.975	0.559
15	302	91	1.000	1.139	0.962	0.438	0.960	0.432	0.965	0.438	0.982	0.513

bootstrap prediction intervals but these were computationally more burdensome and performed very poorly in some non-normal cases; see the Supplementary Material. Finally, we included an interval (Direct) based on the area mean (a direct estimator) and its variance under a homogeneous model for each area (equivalently, under simple random sampling without replacement).

For each simulation setting, we report the size N_i and the sample size n_i for each small area. For each interval, for every small area we computed the empirical coverage probabilities (Cvge) and the average half-length of the interval divided by 1.96 (Alen). Alen is therefore also the average of the root mean squared error estimates.

The full set of results with the average population mean and median across

Table 2: Simulated coverage and length of prediction intervals when α_i has a mixture distribution and e_{ij} has a normal distribution with variances $\dot{\sigma}_\alpha^2 = 64$ and $\dot{\sigma}_e^2 = 100$, respectively.

Method			Direct		sam-LW		sam-RM		clp-LW		clp-PR	
Area	N_i	n_i	Cvge	Alen	Cvge	Alen	Cvge	Alen	Cvge	Alen	Cvge	Alen
1	191	4	0.970	7.049	0.968	4.945	0.910	3.439	0.967	4.945	0.901	3.425
2	61	8	0.983	4.918	0.965	3.294	0.924	2.718	0.965	3.294	0.918	2.791
3	114	8	0.984	5.156	0.961	3.408	0.929	2.759	0.959	3.408	0.921	2.792
4	48	16	0.992	3.150	0.962	2.040	0.945	1.876	0.942	2.040	0.945	2.173
5	40	18	0.993	2.681	0.964	1.747	0.942	1.642	0.940	1.747	0.949	2.075
6	119	36	0.999	2.171	0.954	1.391	0.938	1.329	0.949	1.391	0.956	1.554
7	201	60	0.996	1.684	0.952	1.081	0.947	1.051	0.954	1.081	0.971	1.238
8	209	63	0.998	1.637	0.954	1.053	0.952	1.025	0.952	1.053	0.977	1.211
9	212	64	0.998	1.622	0.953	1.044	0.947	1.017	0.950	1.044	0.978	1.202
10	245	74	0.997	1.509	0.956	0.971	0.948	0.950	0.954	0.971	0.978	1.125
11	279	84	0.996	1.417	0.941	0.912	0.933	0.895	0.939	0.912	0.973	1.061
12	299	90	0.994	1.374	0.947	0.881	0.943	0.866	0.944	0.881	0.972	1.028
13	311	93	0.996	1.354	0.949	0.868	0.945	0.854	0.947	0.868	0.977	1.012
14	346	104	0.997	1.275	0.955	0.820	0.954	0.809	0.960	0.820	0.987	0.961
15	382	115	0.998	1.222	0.951	0.779	0.950	0.771	0.949	0.779	0.978	0.917

populations for each area is given in the Supplementary Material; we present and discuss illustrative cases below. Table 1 shows the empirical coverage and the relative length of the five intervals for the setting with variances $\dot{\sigma}_\alpha^2 = 4$ and $\dot{\sigma}_e^2 = 25$ when α_i and e_{ij} have normal distributions; Table 2 shows the results for the setting with variances $\dot{\sigma}_\alpha^2 = 64$ and $\dot{\sigma}_e^2 = 100$ when α_i has a mixture distribution and e_{ij} has a normal distribution; results for the remaining 48 simulation settings are similar. The areas are presented and labeled in order of increasing sample size. Simulation standard errors for the coverage probabilities can be obtained as $\{\text{Cvge}(1-\text{Cvge})/1000\}^{1/2}$; they are approximately 0.008 or smaller.

The simulation results show that our asymptotic results based on both g and

n_L going to infinity provide useful approximations that work well. The empirical coverages of the intervals are close to the nominal level and tend to the nominal level as g and n_L increase, confirming our large g and n_L asymptotic results. For sample sizes from 4 to 8, sam-LW and clp-LW conservative in coverage and wider than sam-RM (Rao and Molina, 2015) and clp-PR (Prasad and Rao, 1990) which are optimistic and narrower. However, beyond sample size 10, sam-LW and sam-RM perform similarly and effectively; the mean squared error estimator PR typically exceeds LW, making clp-PR more conservative than clp-LW. The simple direct intervals are very conservative and much wider than the other intervals for all sample sizes. So, while the direct intervals can be used in this context, even when the sample size is large, they lose considerable efficiency relative to the model-based intervals.

5. Design-based simulations

We obtained data from the Dairy Survey component of the 2002 Consumer Expenditure Survey conducted by the U.S. Bureau of the Census for the U.S. Bureau of Labor Statistics; the data are available from `https://www.bls.gov/cex/pumd_data.htm`. We treated the consumer expenditure on fresh milk products (MILKPROD) as the survey variable of interest and considered the problem of estimating the average consumer expenditure on fresh milk products in different states (small areas). We used the total expenditure on food (FODTOT), the number of persons under age 18 in the family (PRSLT18) and the total family income before taxes in the last 12 months (FINCBFX) as the auxiliary variables. This data set is similar to that used in Arora and Lahiri (1997); they used the earlier 1989 survey, focussed on the expenditure on fresh whole milk, and potentially used different auxiliary variables. If we knew the means of the auxiliary variables for each state, we could use our methods to estimate the average expenditure on fresh milk products in 2002 in each state. As we do not have this information, we instead treated the data set as a pseudo-population and sampled from it. We repeated this sampling 1000 times, implementing a design-based simulation from our fixed population to evaluate the design-based properties of our proposed model-based methods.

In creating the population, we discarded 6 states with fewer than 10 observations, leaving us with $N = 4022$ observations from $g = 34$ states with between $N_L = 36$ and $N_U = 397$ observations from each state. We centered the auxiliary variables about their area means (adding $cent$ to their variable name) and then included the area means (adding avg to the variable name) as between state variables so that we have $p_b = 3$ plus $p_w = 3$ auxiliary variables. The area means give the average per family of each variable for each state. We selected the 1000 samples independently by simple random sampling without replacement from each state with $n_L = 20$ by setting $n_i = 20$ if $N_i < 50$, $n_i = \lfloor 0.5 * N_i \rfloor$ if $50 \le N_i \le 100$, and $n_i = \lfloor 0.25 * N_i \rfloor$ if $N_i > 100$, where $\lfloor \quad \rfloor$ is the integer part function. In each sample, we fitted the model (5.14)

$$
\begin{aligned}
&MILKPROD_{ij} &(5.14)\\
&= \beta_0 + \beta_1 FODTOTavg_i + \beta_2 PRSLT18avg_i + \beta_3 FINCBFXavg_i\\
&+ \beta_4 FODTOTcent_{ij} + \beta_5 PRSLT18cent_{ij} + \beta_6 FINCBFXcent_{ij} + \alpha_i + e_{ij},
\end{aligned}
$$

using `lmer` and computed the 95% model-based prediction intervals (2.9) and (2.10), as well as sam-RM, clp-PR and the direct interval based on the small area mean. As in the model-based simulation, we report the empirical design-coverage (Cvge) and the relative design-expected length (Rlen) of the prediction intervals; the results over 1000 samples together with the standardised population EBLUPs $\hat{\alpha}_i/\hat{\sigma}_\alpha$ for each state are shown in Table 3. The relative design-bias and the design RMSEs of sam and clp, together with the design-averages of the LW, RM and PR estimators of the RMSEs are available in the Supplementary Material.

To interpret the results, we partition the states into three groups: Group 3 with three states $\{2, 9, 27\}$ for which the design-coverage of all four model-based intervals is well below the nominal level; Group 2 with six states $\{8, 22, 24, 32, 37, 50\}$ for which at least one model-based interval has design-coverage below the nominal level, but not all model-based intervals perform poorly; and Group 1 with the remaining twenty-five states for which the design-coverage of all four model-based intervals is above the nominal level.

For Group 1, the design-coverages are mostly conservative and similar across

Table 3: Simulated design-coverage and design-expected length of nominal 95% confidence intervals for the average consumer expenditure on fresh milk products in each state in 2002. * identifies states in Group 3 and † identifies states in Group 2.

Method				Direct		sam-LW		sam-RM		clp-LW		clp-PR	
STATE	N_i	n_i	$\hat{\alpha}_i/\hat{\sigma}_\alpha$	Cvge	Alen	Cvge	Alen	Cvge	Alen	Cvge	Alen	Cvge	Alen
16	36	20	0.123	0.964	0.383	0.998	0.441	0.992	0.370	1.000	0.441	0.993	0.376
$50^\dagger$	37	20	0.526	0.912	0.677	0.902	0.449	0.825	0.373	0.917	0.449	0.696	0.375
31	42	20	-0.332	0.991	0.405	1.000	0.479	0.994	0.386	1.000	0.479	0.957	0.375
$22^\dagger$	43	20	-0.489	0.908	0.227	1.000	0.484	1.000	0.388	1.000	0.484	0.845	0.375
21	44	20	-0.188	0.998	0.421	1.000	0.489	0.999	0.389	1.000	0.489	0.980	0.375
15	45	20	0.087	0.976	0.414	0.998	0.493	0.996	0.391	1.000	0.493	0.992	0.376
$32^\dagger$	46	20	0.737	0.909	0.637	0.930	0.498	0.815	0.392	0.849	0.498	0.497	0.375
$37^\dagger$	47	20	-0.486	0.999	0.376	1.000	0.502	0.979	0.393	1.000	0.502	0.843	0.375
1	52	26	-0.010	0.993	0.254	1.000	0.410	1.000	0.343	1.000	0.410	0.998	0.362
45	53	26	0.047	0.944	0.454	0.979	0.414	0.936	0.346	0.996	0.414	0.979	0.362
2^*	58	29	1.180	0.866	0.524	0.781	0.389	0.676	0.328	0.395	0.389	0.356	0.355
9^*	59	30	-1.094	0.575	0.281	0.741	0.379	0.655	0.324	0.385	0.379	0.347	0.354
41	65	32	0.128	0.966	0.434	0.993	0.373	0.982	0.317	1.000	0.373	0.996	0.349
49	67	34	0.004	0.987	0.361	0.996	0.356	0.984	0.306	1.000	0.356	0.996	0.345
18	71	36	-0.265	0.951	0.466	0.979	0.346	0.932	0.299	0.998	0.346	0.987	0.341
27^*	76	38	1.317	0.935	0.560	0.765	0.340	0.691	0.294	0.463	0.340	0.474	0.338
$8^\dagger$	82	41	0.666	0.900	0.498	0.866	0.327	0.797	0.285	0.845	0.327	0.826	0.332
13	93	46	-0.088	0.988	0.333	0.992	0.310	0.984	0.273	0.999	0.310	1.000	0.324
$24^\dagger$	94	47	-0.907	0.950	0.285	0.947	0.305	0.893	0.270	0.795	0.305	0.799	0.324
29	98	49	-0.265	0.991	0.249	0.999	0.299	0.999	0.265	1.000	0.299	0.999	0.319
53	99	50	0.242	0.984	0.264	1.000	0.294	0.998	0.262	1.000	0.294	0.998	0.318
55	119	30	-0.462	0.998	0.424	0.997	0.467	0.970	0.354	1.000	0.467	0.940	0.353
51	122	30	0.022	0.994	0.472	1.000	0.469	0.994	0.355	1.000	0.469	0.998	0.353
25	126	32	0.074	0.991	0.396	1.000	0.452	1.000	0.347	1.000	0.452	1.000	0.349
4	133	33	0.133	0.982	0.480	0.997	0.447	0.988	0.344	0.998	0.447	0.995	0.347
26	139	35	-0.325	0.988	0.411	0.998	0.433	0.990	0.337	1.000	0.433	0.989	0.343
34	160	40	-0.160	0.988	0.401	0.996	0.405	0.987	0.323	0.998	0.405	0.996	0.334
17	161	40	-0.301	0.996	0.330	1.000	0.406	0.996	0.323	1.000	0.406	0.997	0.334
39	229	57	-0.155	0.984	0.308	0.997	0.340	0.992	0.287	1.000	0.340	1.000	0.307
42	261	65	-0.538	0.993	0.327	0.977	0.318	0.945	0.274	0.983	0.318	0.952	0.297
36	280	70	0.019	0.974	0.362	0.978	0.306	0.961	0.267	0.985	0.306	0.984	0.291
12	283	71	0.594	0.946	0.321	0.977	0.304	0.947	0.265	0.978	0.304	0.949	0.290
48	305	76	-0.430	0.990	0.286	0.983	0.294	0.963	0.259	0.989	0.294	0.985	0.284
6	397	99	0.595	0.975	0.291	0.966	0.258	0.951	0.235	0.966	0.258	0.965	0.262

all intervals. The design-biases for sam are smaller than for clp except in state 49 where the design-biases are very similar. There is no simple relationship between the design-average of the direct variance estimates and the design-averaged LW, RM or PR estimates. The design-averaged LW estimator is larger than the design-averaged RM estimator; the design-averaged LW estimator is larger than the design-averaged PR estimator in all except 4 states (6, 13 , 29 and 53) where PR has a slightly smaller design-average than LW.

Group 3 comprises the three states for which all four model-based intervals (sam-LW, sam-RM, clp-LW and clp-PR) have design-coverage well below the nominal level. The direct intervals perform relatively well for states 2 and 27, but poorly for state 9. It turns out that state 9 has at least one large outlier and the accuracy of the predictions depends heavily on whether the outlier is included in the sample or not; failure to include the outlier leads to direct intervals with poor design-coverage, lowering the overall design-coverage of the intervals. The outlier also affects the the model-based intervals. Among the model-based intervals, sam-LW has the best design-coverage, followed by sam-RM, and then clp-PR and clp-LW. Both point estimates sam and clp are design-biased, clp more than sam. The design-averaged RMSE estimator LW is greater than the design-averaged RM (or PR), explaining why sam-LW has better design-coverage than sam-RM.

Group 2 comprises the states for which at least one but not all four model-based intervals (sam-LW, sam-RM, clp-LW and clp-PR) have design-coverage be-

Table 4: Parameter estimates (REML) for modelling the consumer expenditure on fresh milk products population.

Effect	β_0	β_1	β_2	β_3	β_4	β_5	β_6	σ_α	σ_e
Estimate	2.623	1.163	0.924	0.016	4.793	0.710	0.002	0.176	8.560
Std Error	0.762	5.087	0.553	0.008	0.366	0.040	0.001		

NOTE: lmer does not compute standard errors for the variance components.

low the nominal level. Usually, sam-LW has the best design-coverage of the model-based intervals (although it has below nominal design-coverage in state 8) and similar design-coverage to the direct intervals. The sam-LW intervals are narrower than the direct intervals in states 50, 32 and 8 while the direct intervals are narrower than sam-LW in states 22, 37 and 24. Usually clp-PR has the lowest design-coverage, but clp-LW has lowest design-coverage in state 24 and sam-RM has the lowest design-coverage in state 8. Clp has larger design-bias than sam for Group 2.

These results show that the sam-LW intervals generally have better design-based properties than the clp intervals in our consumer expenditure on fresh milk products population. Nonetheless, we did not expect to see results like those in Group 3 so we explored why these results occur.

As we have access to the whole population (which is unusual in practice), we are able to explore the population. We used `lmer` from the R package `lme4` to fit the nested error regression model (5.14) to the population data. Figure 1 shows a normal QQ-plot of the EBLUPs (with approximate 95% prediction intervals com-

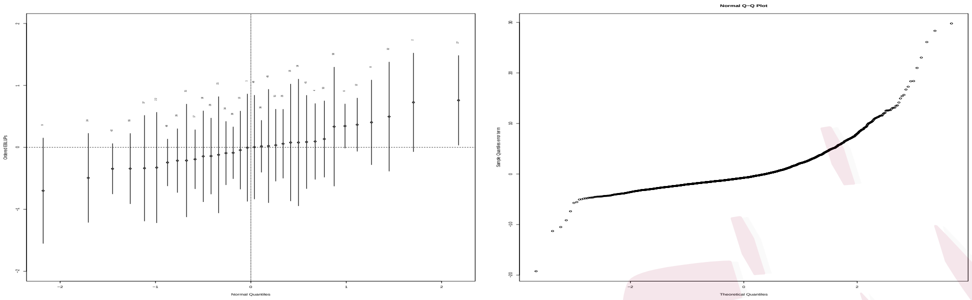

Figure 1: Normal QQ-plots for the EBLUPs $\hat{\alpha}_i$ and estimated errors $\hat{e}_{ij}$ from the model (5.14) fitted to the consumer expenditure on fresh milk products in 2002 data.

puted as in Lyu and Welsh (2022a)) and a normal QQ-plot of the errors from the fitted model. Figure 1 shows that it is plausible to treat the random effects as approximately normally distributed, but the errors have an asymmetric long-tailed distribution and it is less plausible to treat the errors as normally distributed. In a model-based analysis, rather than simply relying on the asymptotic theory, we could consider using transformations to improve the fit, make predictions on the transformed scale and then invert the prediction intervals (possibly adjusting for back-transformation bias). However, for our design-based analysis, as is arguably usual in practice, we keep the data on the raw scale. Table 4 shows the parameter estimates and standard errors for the fitted model. Although some coefficients are not significant, we retain them in the model. We see that $\hat{\sigma}_e^2/\hat{\sigma}_a^2 = 48.72$ is large so that the within area correlation is very small (approximately 0.02).

The normal QQ-plot of the EBLUPs in Figure 1 shows that the Group 3 states have extreme EBLUPs and the Group 2 states have EBLUPs in the tails of the distribution but these are mixed in with some EBLUPs from Group 1 states. This mixing suggests that the EBLUPS alone do not identify the group to which a state belongs. Another potentially important value is the sample size n_i in each state. Figure 2 shows the population standardised EBLUPS $\hat{\alpha}_i/\hat{\sigma}_\alpha$ plotted against the sample size n_i for each state; these variables are also included in Table 3. States plotted in the top and bottom left of the plot (high standardised EBLUPs and small to moderate sample size) are in Groups 2 and 3; states at the top or bottom right (small standardised EBLUPs or large standardised EBLUPS with large sample sizes) are in Group 1 with all the others states. Specifically, states 6, 12 and 42 have relatively extreme EBLUPs but larger sample sizes so are in Group 1. This suggests that both the magnitude of the standardised EBLUPs and the sample size determine the difficulty of estimating a particular state.

We cannot rule out the possibility that the results of the design-based simulation using the consumer expenditure on fresh milk products data are at least in part due to failures of the model (5.14). However, we can explore whether similar results occur when the model is correct by carrying out a set of additional design-based simulations. We used the same 48 settings as in our model-based simulation, but we generated a single population for each setting and then selected 1000 samples from it by simple random sampling without replacement. We fitted the

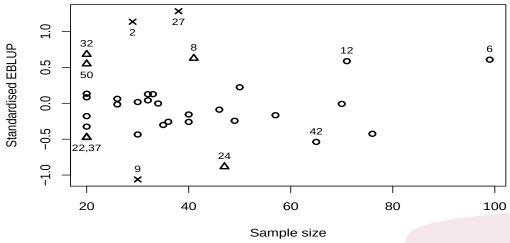

Figure 2: Plot of the population standardised EBLUPS $\hat{\alpha}_i/\hat{\sigma}_\alpha$ against sample size in each state for the consumer expenditure on fresh milk products. Group 1 states are plotted as circles, Group 2 as triangles and Group 3 as crosses. Selected states are labelled by state number.

model and computed the same 95% prediction intervals as before and evaluated their design-based properties. The full set of results are available in the Supplementary Material. The overall conclusion is that when $\dot{\sigma}_e^2/\dot{\sigma}_\alpha^2$ is large, areas with extreme EBLUPs and small to moderate sample sizes are difficult to estimate well in the design-based framework.

Why does the size of the random effect for an area matter in the design-based framework but not in the model-based framework? In the model-based framework, the population and hence the random effect for an area is generated anew for each replication. This means that in each sample we are estimating a realisation of an independent and identically distributed random variable so, over model-based replications, we are estimating the expected value of the random variable which is zero under the model. In the design-based framework, the population is fixed and the replication is over independent samples from this fixed population. Once generated, the random effects are fixed so, in areas with extreme random effects, the EBLUPs are estimating the expected values of extreme order statistics which are not zero and difficult to estimate. This flows through into estimating the area mean of the survey variable for areas with large random effects. In our design-based simulation, we used the population EBLUPs to assess the difficulty in estimation, but in practice, as shown in Lyu and Welsh (2022a), we would use the sample EBLUPs to estimate the population random effects.

The above discussion suggests that when we want to achieve good (model-assisted) design-based rather than model-based performance, we should treat the random effects in the model as fixed. To check this intuition (and indirectly confirm the argument above), we repeated our design-based simulations treating α_i as fixed and examined the empirical design coverage and the relative design-expected length of the model-based prediction intervals constructed under the fixed area effects model. Details and results are included in the Supplementary Material. The design-coverage results for both composite and synthetic methods are generally good, confirming our intuition.

6. Discussion

In this paper, we considered model-based small area estimation under the nested error regression model. We discussed two targets of interest, the small area means and the conditional linear predictors of the small area means, and the construction of mixed model estimators (EBLUPs) of these two targets. We established asymptotic linearity results and central limit theorems for these estimators which allow us to establish asymptotic equivalences between estimators, to approximate their sampling distributions, obtain simple expressions for and construct simple estimators of their asymptotic mean squared errors, and justify asymptotic prediction intervals. Our new results are established under the asymptotic framework of increasing numbers of small areas and increasing numbers of units in each area; we report model-based simulations that show that these results are applicable in quite small, finite samples, establishing that they fill important theoretical gaps and are useful in practice. In particular, our mean squared error estimator performs as well or better than the widely-used Prasad and Rao (1990) and Rao and Molina (2015) estimator and is much simpler, so it is easier to interpret and consequently provides more insight. We also carried out a design-based simulations using real data on consumer expenditure on fresh milk products. This simulation produced some surprising results which we explained and interpreted through analysis of the population and further design-based simulations. The simulations together highlight under-appreciated differences between the model- and design-based properties of mixed model estimators in small area estimation.

Following the suggestion of a referee, we also applied our proposed method to the Iowa corn data presented in Battese et al. (1988). This is a challenging data set for both fixed and increasing area size methods because there are only 12 areas, three of which have sample size 1, and the largest sample size is 6. Details of our analysis are reported in the Supplementary Material. We observe that the sam and clp predictors are very similar in this example. The LW estimator of the MSE proposed in this study is larger than both the RM and PR estimators when the sample size is very small, specifically in the range from 1 to 4. Nevertheless, the difference between the proposed LW estimator and both RM and PR estimators tends to decrease with increasing sample size. Notably, when the sample size reaches 6, the proposed LW estimator is smaller than both the RM and PR estimators. There is no simple pattern in the relationship between the sample mean and the sam/clp predictors or between the variance of the sample mean and the other MSE estimators. With only a single sample and no knowledge of the true values, it is difficult to reach any conclusions about the validity of the methods in this application.

Given the extensive literature on small area estimation, it is important to acknowledge that in this paper we have considered only one of many interesting and important scenarios. Future work should include applying the asymptotic approach to area level models, outlier robust estimators and the extensions to the basic nested error regression model.

Acknowledgement

We are grateful to the Associate Editor and two anonymous referees for very helpful comments that improved the presentation of our paper. The work of the second author was supported by Australian Research Council Discovery Project DP230101908.

References

Arora, V. and P. Lahiri (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica 7*(4), 1053–1063.

Battese, G. E., R. M. Harter, and W. A. Fuller (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.

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

Costa, Ã., A. Satorra, and E. Ventura (2003, Dec). Using composite estimators to improve both domain and total area estimation. Economics working papers, Department of Economics and Business, Universitat Pompeu Fabra.

Das, K., J. Jiang, and J. Rao (2004). Mean squared error of empirical predictor. *The Annals of Statistics 32*(2), 818–840.

Datta, G. S. and P. Lahiri (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica 10*(2), 613–627.

Fay, R. and R. Herriot (1979). Estimates of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association 74*, 269–277.

Haslett, S. and A. Welsh (2019). EBLUPs: Empirical best linear unbiased predictors. *Wiley StatsRef: Statistics Reference Online*.

Jiang, J. and P. Lahiri (2006a). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association 101*(473), 301–311.

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

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

Jiang, J., T. Nguyen, and J. S. Rao (2011). Best predictive small area estimation. *Journal of the American Statistical Association 106*(494), 732–745.

Kackar, R. N. and D. A. Harville (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics-theory and Methods 10*, 1249–1261.

Lahiri, P. and J. Rao (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association 90*(430), 758–766.

Lehtonen, R. and A. Veijanen (2009). Design-based methods of estimation for domains and small areas. In *Handbook of Statistics*, Volume 29, pp. 219–249. Elsevier.

Lyu, Z. and A. Welsh (2022a). Asymptotics for EBLUPs: Nested error regression models. *Journal of the American Statistical Association 117*, 2028–2042.

Lyu, Z. and A. Welsh (2022b). Increasing cluster size asymptotics for nested error regression models. *Journal* *of Statistical Planning and Inference 217*, 52–68.

Morales, D., M. D. Esteban, A. PÃľrez, and T. Hobza (2021). *A Course on Small Area Estimation and Mixed Models: Methods, Theory and Applications in R.* Cham: Springer.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science 28*(1), 40–68.

Prasad, N. and J. Rao (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.

Rao, J. (2005). Inferential issues in small area estimation: Some new developments. *Statistics in Transition 7*(3), 513–526.

Rao, J. (2008). Some methods for small area estimation. *Rivista Internazionale di Scienze Sociali 116*(4), 387–406.

Rao, J. and I. Molina (2015). *Small Area Estimation.* John Wiley & Sons.

Saei, A. and R. Chambers (2003a). Small area estimation: A review of methods based on the application of mixed models. *Southampton Statistical Sciences Research Institute Methodology Working Paper M03/16, University of Southampton.*

Saei, A. and R. Chambers (2003b). Small area estimation under linear and generalized linear mixed models with time and area effects. *Southampton Statistical Sciences Research Institute Methodology Working Paper M03/15, University of Southampton.*

Sugasawa, S. and T. Kubokawa (2020). Small area estimation with mixed models: A review. *Japanese Journal of Statistics and Data Science 3*(2), 693–720.

Torabi, M. and J. Rao (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.

Tzavidis, N., S. Marchetti, and R. Chambers (2010). Robust estimation of small-area means and quantiles. *Australian & New Zealand Journal of Statistics 52*(2), 167–186.

Yoon, H.-J. and A. H. Welsh (2020). On the effect of ignoring correlation in the covariates when fitting linear mixed models. *Journal of Statistical Planning and Inference 204*, 18–34.

ZÄĚdÅĆo, T. (2009). On MSE of EBLUP. *Statistical Papers 50*, 101–118.

Ziyang Lyu

E-mail: ziyang.lyu@unsw.edu.au

A.H.Welsh

E-mail: alan.welsh@anu.edu.au