University of Wollongong
Research Online

Centre for Statistical & Survey Methodology
Working Paper Series

Faculty of Engineering and Information Sciences

2011

What Level of Statistical Model Should We Use in Small Domain Estimation?

Mohammad-Reza Namazi-Rad
University of Wollongong, mrad@uow.edu.au

David Steel
University of Wollongong

Recommended Citation

Namazi-Rad, Mohammad-Reza and Steel, David, What Level of Statistical Model Should We Use in Small Domain Estimation?, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 02-11, 2011, 28p.
http://ro.uow.edu.au/cssmwp/52

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

02-11

What Level of Statistical Model Should We Use in Small Domain Estimation?

Mohammad-Reza Namazi-Rad and David Steel

Copyright © 2008 by the Centre for Statistical & Survey Methodology, UOW. Work in progress, no part of this paper may be reproduced without permission from the Centre.

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

What Level of Statistical Model Should We Use in Small Domain Estimation?

Mohammad-Reza Namazi-Rad

Centre for Statistical and Survey Methodology University of Wollongong, NSW 2522, Australia

mohammad_ namazi@uow.edu.au

David Steel

Centre for Statistical and Survey Methodology University of Wollongong, NSW 2522, Australia

david_ steel@uow.edu.au

May 2011

Abstract

If unit-level data are available, Small Area Estimation (SAE) is usually based on models formulated at the unit level, but they are ultimately used to produce estimates at the area level and thus involve area-level inferences. This paper investigates the circumstances when using an area-level model may be more effective. Linear mixed models fitted using different levels of data are applied in SAE to calculate synthetic estimators and Empirical Best Linear Unbiased Predictors (EBLUPs). The performance of area-level models is compared with unit-level models when both individual and aggregate data are available. A key factor is whether there are substantial contextual effects. Ignoring these effects in unit-level working models can cause biased estimates of regression parameters which is referred to as the ecological fallacy. The contextual effects can be automatically accounted for in the area-level models. Using synthetic and EBLUP techniques, small area estimates based on different levels of linear mixed models are studied in a simulation study.

Keywords: *Contextual Effect; EBLUP; Ecological Fallacy; Small Area Estimation; Synthetic Estimator.*

1 Introduction

There are increasing demands for comprehensive statistical information not only at national levels but also for sub-national domains in many countries. Statistical Bureaus and survey organizations are using sample surveys to produce estimates for the total population and possibly large regions. However, there are often difficulties in producing useful and reliable estimates for various local areas and other small domains using standard estimation methods due to small sample sizes. Some areas may have no sample at all.

Small area estimation (SAE) involves techniques based on statistical models to produce estimates for relatively small geographic sub-populations such as cities, provinces or states, for which the available survey data does not allow the calculation of reliable direct estimates. Usually auxiliary variables related to the target variable are used in statistical models to calculate the required estimates in different SAE techniques (Rao, 2003). A key feature of this approach is that the statistical model used does not involve area-specific parameters and estimation of the parameters can use data from the entire sample. These parameter estimates are then used with population information about auxiliary variables for each area to produce small area estimates.

A wide variety of estimation methods have been developed to handle SAE problems. Initially, demographic and design-based methods were used, but more sophisticated model-based methods have been increasingly employed over the last two decades (Khoshgooyanfard and Taheri Monazah, 2006). See Rao (2003) and Longford (2005) for comprehensive discussions on different SAE methods.

Statistical models for small area estimation purposes can be formulated at the individual or aggregated levels. When sufficient information about the geographic indicators for target areas are available for all individuals in the sample, the usual approach is to estimate regression coefficients and variance components based on a unit-level linear model. However, it is also possible to aggregate the data to area level and estimate these parameters based on a linear model for the area means. When the unit-level model is properly specified, the parameter estimates from the individual and aggregated level analysis will have the same expectation but we would expect that parameter estimates obtained using unit-level data to have less variance. However, in practice the parameter estimates from

different levels of data analysis often differ due to some model misspecifications. Given that the targets of inference are at the area-level, the question arises as to whether it is sometimes preferable to use an area-level analysis and under what conditions an area-level analysis may be better. In practice, if the correct population model includes the contextual effect of the area-level means, the area-level analysis should produce less biased estimates of the regression coefficients.

The main purpose of this paper is to evaluate unit-level and area-level modeling approaches when both individual-level and aggregate data are available. Using a Mont-Carlo simulation, parameter estimates based on different levels of statistical modeling are studied when area-level means are involved in the unit-level population model as contextual effects. In this study, the estimators will be calculated based on synthetic and Empirical Best Linear Unbiased Predictor (EBLUP) methods. The effects of these methods on the efficiency of small area estimates are also evaluated.

2 Linear Mixed Models in Small Area Estimation

Indirect techniques for SAE purposes mostly rely on statistical models which borrow strengths with an explanation of possible relations to other auxiliary data recourses. Efficient models to this extent usually include random effects to explain the variations between target areas within the population as well as several covariates for available auxiliary variables (Chambers and Tzavidis, 2006). As mentioned before, statistical models utilized for SAE purposes can be unit-level or area-level.

2.1 Unit- and Area-level Population Models

Consider a population of size N divided into K small areas with N_k individuals in the kth small area ($N = \sum_{k=1}^{K} N_k$). A unit-level mixed linear model which relates the unit population values of the study variable to unit-specific auxiliary variables including both fixed and random effects is:

$$\begin{gathered} Y_{ik} = \mathbf{X}'_{ik}\beta + u_k + e_{ik} \ ; \quad i = 1, \ldots, N_k \ \ \& \ \ k = 1, \ldots, K \\ u_k \overset{iid}{\sim} N(0, \sigma_u^2) \ \ ; \ \ e_{ik} \overset{iid}{\sim} N(0, \sigma_e^2) \end{gathered} \tag{1}$$

where $\mathbf{X}'_{ik} = [1\ X_{ik1}\ \ldots\ X_{ikP}]$ is a vector of P auxiliary variables for ith unit within the kth area and $\beta' = [\beta_0\ \beta_1\ \ldots\ \ \beta_P]$ denotes the vector of unknown regression parameters. The random effect for the kth area is denoted by u_k and e_{ik} is the random error for the ith individual within the kth area. The random effects and random errors are independently distributed in the model.

Area-level models can be derived from the unit-level model by aggregating or averaging the data to area levels. A standard area-level linear mixed model obtained from (1) for the population area means is given as:

$$\begin{aligned} \bar{Y}_k &= \bar{\mathbf{X}}'_k\beta + u_k + \bar{e}_k \ \ ; \ \ k = 1, \ldots, K \\ u_k &\overset{iid}{\sim} N(0, \sigma_u^2) \ \ ; \ \ \bar{e}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} e_{ik} \sim N(0, \frac{\sigma_e^2}{N_k}) \end{aligned} \tag{2}$$

where $\bar{\mathbf{X}}'_k = [1\ \bar{X}_{k1}\ \ldots\ \bar{X}_{kP}]$ is the vector of population mean values for the P auxiliary variables within the kth area.

The linear mixed models used in SAE relate the unit (or area) values of the study variable to P unit-specific (or area-specific) auxiliary variables within the target population can also be presented in matrix forms as follows:

$$\begin{aligned} \textbf{Unit-Level Population Model}: \mathbf{Y} &= \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \\ \mathbf{u} \sim N(\mathbf{0},\ \sigma_u^2\ \mathbf{I}_K) \ \ &; \ \ \mathbf{e} \sim N(\mathbf{0},\ \sigma_e^2\ \mathbf{I}_N) \end{aligned} \tag{3}$$

$$\begin{aligned} \textbf{Area-Level Population Model}: \bar{\mathbf{Y}} &= \bar{\mathbf{X}}\beta + \mathbf{u} + \bar{\mathbf{e}} \\ \mathbf{u} \sim N(\mathbf{0},\ \sigma_u^2\ \mathbf{I}_K) \ \ &; \ \ \bar{\mathbf{e}} \sim N\big(\mathbf{0},\ diag(\tfrac{\sigma_e^2}{N_1},\ \ldots,\ \tfrac{\sigma_e^2}{N_K})\big). \end{aligned} \tag{4}$$

Here, $\mathbf{Y}$ and $\mathbf{e}$ are column vectors with N elements, $\bar{\mathbf{Y}}$ and $\bar{\mathbf{e}}$ are column vectors with K elements, $\mathbf{X}$ and $\bar{\mathbf{X}}$ are respectively $N \times (P+1)$ dimensional and $K \times (P+1)$ dimensional matrices. β and $\mathbf{u}$ are two column vectors with $(P+1)$ and K elements, respectively. Finally, $\mathbf{Z}$ is a $N \times K$ dimensional matrix that includes 1s and 0s which assigns the same value of u_k to all the rows referring to the units within the kth area. Note that, matrices are shown by bold print in this paper.

A basic area-level model seems appropriate when the data are available just at the area level and the estimation process is possible only based on aggregate data. We will

consider the issue of whether there are advantages in using an area-level model when the individual-level data is available, given that the final small area estimates are produced at the area level.

2.2 Parameter Estimation using Unit-level Data

Sample surveys allow estimation and inference about a large population when the resources available do not permit collecting relevant information from every member of the target population. In this paper, sample s of size n is assumed to be selected from the target population U. The part of the whole sample s which falls into the kth area is $s_k = s \bigcap U_k$ and is of size n_k.

It is often the case that reliable direct estimates can not be obtained based on the available sample data due to small sample sizes in all or some of the areas. In order to calculate model-based estimators, a model should be developed to specify the relationship between the auxiliary information and variable of interest based on the available sample data. In this paper, the term *working model* is used for the statistical model to be fitted on the sample data and *population model* for the correct model assumed for the population data. The working model may not be correct in practice.

A simple unit-level working model which can be fitted on individual-level sample data is given as:

$$\begin{gathered} \mathbf{y} = \mathbf{x}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \\ \mathbf{u} \sim N(\mathbf{0},\ \sigma_u^2\ \mathbf{I}_K)\ \ ;\ \ \mathbf{e} \sim N(\mathbf{0},\ \sigma_e^2\ \mathbf{I}_n) \end{gathered} \tag{5}$$

It will be noted that, lowercase letters refer to sample statistics and uppercase to population statistics. Hence, $\mathbf{y}$ is a vector which contains sample values for the target variable and $\mathbf{x}$ denotes the matrix of auxiliary data values for the individuals falling into the sample. The corresponding data for s_k are $\mathbf{y}_k$ and $\mathbf{x}_k$ $(k = 1\ ,2\ ,\ldots,K)$. We assume that the sampling scheme used is uninformative. Therefore, the same model can be used for the sample and population at the individual level.

Usually the model parameter estimates are calculated using the information obtained from the sample surveys. In order to define the Maximum Likelihood (ML) technique for a simple random sample design, $L(\mathbf{y}\ ;\beta, \sigma_u^2, \sigma_e^2)$ is assumed to be the twice differentiable

probability density function for variable $\mathbf{y}$,

$$L(\mathbf{y}\ ;\beta,\sigma_u^2,\sigma_e^2) = c\ |\Sigma|^{-\frac{1}{2}}\ exp\big[-\frac{1}{2}\ (\mathbf{y}-\mathbf{x}\beta)'\ \Sigma^{-1}(\mathbf{y}-\mathbf{x}\beta)\big] \tag{6}$$

where c is a constant value and Σ is the block-diagonal variance-covariance matrix as follows:

$$\Sigma = diag(\Sigma_k) \tag{7}$$

where:

$$\begin{aligned} &\Sigma_k = \sigma_u^2\mathbf{J}_{n_k} + \sigma_e^2\mathbf{I}_{n_k} \\ &\mathbf{J}_{n_k} = \mathbf{1}_{n_k}\mathbf{1}'_{n_k} \qquad\qquad ;\quad k = 1\ ,2\ ,\ldots,K. \end{aligned} \tag{8}$$

Let $l(\beta,\sigma_u^2,\sigma_e^2;\ \mathbf{y})$ to be the log-likelihood function shown as:

$$\begin{aligned} l(\beta,\sigma_u^2,\sigma_e^2;\ \mathbf{y}) &= ln\big[L(\mathbf{y}\ ;\beta,\sigma_u^2,\sigma_e^2)\big] \\ &= ln(c) - \frac{1}{2}\ ln|\Sigma| - \frac{1}{2}(\mathbf{y}-\mathbf{x}\beta)'\ \Sigma^{-1}(\mathbf{y}-\mathbf{x}\beta) \\ &= ln(c) - \frac{1}{2}\sum_{k=1}^{K} ln|\Sigma_k| - \frac{1}{2}\sum_{k=1}^{K}\varsigma_k'\ \Sigma_k^{-1}\varsigma_k \end{aligned} \tag{9}$$

where:

$$\Sigma_k^{-1} = \sigma_e^{-2}(\mathbf{I}_{n_k} - \tfrac{\gamma_k}{n_k}\ \mathbf{1}_{n_k}\ \mathbf{1}'_{n_k}) \tag{10}$$

in which:

$$\gamma_k = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{n_k}} \qquad \& \qquad \varsigma_k = \mathbf{y} - \mathbf{x}\beta\ . \tag{11}$$

The ML estimates are then calculated by maximizing the right-hand side of the log-likelihood equations (Ruppert *et. al.*, 2003). Assuming σ_u and σ_e to be known, the ML estimator for β is:

$$\hat{\beta}^U = (\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}\mathbf{x}'\Sigma^{-1}\mathbf{y} \tag{12}$$

where $\hat{\beta}^U$ denotes the ML estimated value for the parameter vector β using the unit-level sample data.

Calculating parameter estimates is more challenging when we drop the unrealistic

assumption that variance components are already known. On substitution of $\hat{\beta}^U$ into the log-likelihood expression, the profile log-likelihood function for (σ_e^2 , σ_u^2) can be obtained as follows:

$$l_P(\sigma_u^2, \sigma_e^2) = ln(c) - \frac{1}{2}\, ln|\Sigma| - \frac{1}{2}\, \mathbf{y}'\Sigma^{-1}\big[I - \mathbf{x}(\mathbf{x}\Sigma^{-1}\mathbf{x})^{-1}\mathbf{x}'\Sigma^{-1}\big]\mathbf{y}\ . \tag{13}$$

As there is no closed form solution for maximizing profile-likelihood over (σ_e^2 , σ_u^2), numerical methods are developed. The Fisher scoring algorithm is a form of Newton's method commonly used to find ML parameter estimates in mixed models (Osborne, 1992). The parameters β, σ_e^2 and σ_u^2 can be estimated by Fisher scoring algorithm. Alternatively, mixed model packages use Restricted Maximum Likelihood (REML) estimation techniques in order to maximize the restricted log-likelihood expression and estimate the variance parameters. The restricted log-likelihood is:

$$l_R(\sigma_u^2, \sigma_e^2) = l_P(\sigma_u^2, \sigma_e^2) - \frac{1}{2} log|\mathbf{x}\Sigma^{-1}\mathbf{x}|\ . \tag{14}$$

The additional term in the equation for the restricted log-likelihood (l_R) is based on contrast arguments that account for estimation of the β (McCulloch *et. al.*, 2008). Detailed discussions about different methods of estimating model parameters can be found in Ruppert *et al.* (2003). ML and REML techniques are the most common strategies being used for calculating model parameter estimates. Here, an estimation technique is presented using Fisher scoring algorithm for ML estimation.

Longford (1993) defined the Fisher scoring algorithm for estimating a value for parameter θ as follows:

$$\theta_{(t+1)} = \theta_{(t)} + \mathcal{I}^{-1}(\theta_{(t)})\ \mathcal{S}(\theta_{(t)}) \tag{15}$$

where:

$$\mathcal{I}(\theta^*) = -E\Big(\frac{\partial^2 l}{\partial\theta^*\partial\theta^{*\prime}}\Big) \quad \& \quad \mathcal{S}(\theta^*) = \frac{\partial l}{\partial\theta}\bigg|_{\theta=\theta^*} \tag{16}$$

The notations (t) and (t+1) denote the previous and new estimated values for these parameters, respectively. In order to use the Fisher's scoring algorithm for σ_u^2 and σ_e^2, λ is defined to be the variance ratio ($\lambda = \sigma_u^2/\sigma_e^2$). Then, the estimated value for this parameter

can be calculated numerically, as below: [Longford, 1993; p.108]

$$\frac{\partial l(\theta^*;\ \mathbf{y})}{\partial \lambda} = -\frac{1}{2}\sum_{k=1}^{K} \mathbf{1}'_{n_k}\mathbf{W}_k^{-1}\mathbf{1}_{n_k} + \frac{1}{2\sigma_e^2}\sum_{k=1}^{K}\left(\varsigma'_k\mathbf{W}_k^{-1}\mathbf{1}_{n_k}\right)^2 \tag{17}$$

and,

$$\begin{aligned} -E\Big(\frac{\partial^2 l(\theta^*;\ \mathbf{y})}{\partial^2 \lambda}\Big) &= \frac{1}{2}\sum_{k=1}^{K}\left(\mathbf{1}'_{n_k}\mathbf{W}_k^{-1}\mathbf{1}_{n_k}\right)^2 = \frac{1}{2}\sum_{k=1}^{K}\left(f_k^{-1}\mathbf{1}'_{n_k}\mathbf{1}_{n_k}\right)^2 \\ -E\Big(\frac{\partial^2 l(\theta^*;\ \mathbf{y})}{\partial\beta\ \partial\lambda}\Big) &= \mathbf{x}'\frac{\partial \mathbf{W}^{-1}}{\partial\lambda}\ E(e_{ik}) = 0 \end{aligned} \tag{18}$$

where $\theta^* = (\beta,\ \sigma_u^2,\ \sigma_e^2)$, $f_k = 1 + n_k\lambda$ and

$$\begin{aligned} \mathbf{W} = \sigma_e^{-2}\Sigma \quad &; \quad \mathbf{W}_k = \sigma_e^{-2}\big(\sigma_u^2\ \mathbf{1}_{n_k}\ \mathbf{1}'_{n_k} + \sigma_e^2\ \mathbf{I}_{n_k}\big) = \lambda\ \mathbf{1}_{n_k}\mathbf{1}'_{n_k} + \mathbf{I}_{n_k} \\ \mathbf{W}^{-1} = \sigma_e^2\ \Sigma^{-1} \quad &; \quad \mathbf{W}_k^{-1} = \frac{-\sigma_u^2}{\sigma_e^2 + n_k\sigma_u^2}\ \mathbf{1}_{n_k}\ \mathbf{1}'_{n_k} + \mathbf{I}_{n_K}\ . \end{aligned} \tag{19}$$

Then, given estimates $\hat{\beta}^U_{(t)}$ and $\sigma^2_{e(t)}$ of β and σ_e^2, respectively, the new estimated value for the parameter λ can be calculated as follows:

$$\begin{aligned} \hat{\lambda}_{(t+1)} &= \hat{\lambda}_{(t)} + \left[\frac{1}{2}\sum_{k=1}^{K}(f_{k(t)}^{-1}\mathbf{1}'_{n_k}\mathbf{1}_{n_k})^2\right]^{-1}\left[-\frac{1}{2}\sum_{k=1}^{K}(f_{k(t)}^{-1}\mathbf{1}'_{n_k}\mathbf{1}_{n_k}) + \frac{1}{2\hat{\sigma}^2_{e(t)}}\sum_{k=1}^{K}(f_{k(t)}^{-1}\hat{\varsigma}'_{k(t)}\mathbf{1}_{n_k})^2\right] \\ &= \hat{\lambda}_{(t)} + \left[\frac{1}{2}\sum_{k=1}^{K}\frac{n_k^2}{f_{k(t)}^2}\right]^{-1}\left[-\frac{1}{2}\sum_{k=1}^{K}(\frac{n_k}{f_{k(t)}}) + \frac{1}{2\hat{\sigma}^2_{e(t)}}\sum_{k=1}^{K}(f_{k(t)}^{-1}\ \hat{\varsigma}'_{k(t)}\ \mathbf{1}_{n_k})^2\right] \end{aligned} \tag{20}$$

where $f_{k(t)} = 1 + n_k\lambda_{(t)}$, and $\hat{\varsigma}_{k(t)} = \mathbf{y}_k - \mathbf{x}'_k\hat{\beta}^U_{(t)}$.

Given the estimates of β and σ_e^2, the sample data only affect the calculation in equation (20) through $\hat{\varsigma}'_{k(t)}\mathbf{1}_{n_k} = n_k(\bar{y}_k - \bar{\mathbf{x}}'_k\hat{\beta}^U_{(t)})$, which are the area-level residuals. To use the Fisher algorithm for a unit-level mixed linear model, separate consecutive steps are:

- First the parameter β should be estimated based on the Ordinary Least Squares (OLS) method. Then, the initial estimated value for β is given by :

$$\hat{\beta}^U_{(1)} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$$

  where there is no need to estimate Σ.

- Using this initial value, the individual-level residuals can be calculated via:

$$\hat{\mathbf{e}}_{(1)} = \mathbf{y} - \mathbf{x}\hat{\beta}^{U}_{(1)}.$$

- With the use of these residuals, σ_e^2 can be estimated as below:

$$\hat{\sigma}^2_{e(1)} = \frac{1}{n-p}(\mathbf{y} - \mathbf{x}\hat{\beta}^{U}_{(1)})'(\mathbf{y} - \mathbf{x}\hat{\beta}^{U}_{(1)}) \ .$$

- Suppose $\hat{\lambda}_{(t)} = \hat{\sigma}^2_{u(t)}/\hat{\sigma^2_{e(t)}}$ and $f_{k(t)} = 1 + n_k\hat{\lambda}_{(t)}$, then a new estimated value for λ can be calculated through the equation given below. Note that, $\hat{\sigma}_{e(1)}/1000$ is taken to be the initial value for $\hat{\sigma}^2_u$. Therefore, $\lambda_{(1)} = 0.001$, and the scoring function for calculating further values for this parameter is give as:

$$\hat{\lambda}_{(t+1)} = \hat{\lambda}_{(t)} + \Big(\frac{1}{2}\sum_{k=1}^{K}\frac{n_k^2}{f^2_{k(t)}}\Big)^{-1}\Big(-\frac{1}{2}\sum_{k=1}^{K}(\frac{n_k}{f_{k(t)}}) + \frac{1}{2\hat{\sigma}^2_{e(t)}}\sum_{k=1}^{K}(f^{-1}_{k(t)}\ \hat{\varsigma}'_{k(t)}\ \mathbf{1}_{n_k})^2\Big)$$

- Then, $\hat{\sigma}^2_{u(t+1)} = \hat{\lambda}_{(t+1)}\hat{\sigma}^2_{e(t)}$.

- Using $\hat{\sigma}^2_{u(t+1)}$and $\hat{\sigma}^2_{e(t)}$, $\hat{\Sigma}_{(t+1)}$ can be derived based on equation (8), and the new estimate of parameter β is:

$$\hat{\beta}^{U}_{(t+1)} = (\mathbf{x}'\hat{\Sigma}^{-1}_{(t+1)}\mathbf{x})^{-1}\mathbf{x}'\hat{\Sigma}^{-1}_{(t+1)}\mathbf{y} \ .$$

- Now, $\hat{\sigma}^2_{e(t+1)}$ can be calculated by:

$$\hat{\sigma}^2_{e(t+1)} = \frac{1}{n}\ (\mathbf{y} - \mathbf{x}\beta_{(t+1)})'\ \widehat{\mathbf{W}}^{-1}_{(t+1)}\ (\mathbf{y} - \mathbf{x}\beta_{(t+1)}) \ .$$

The steps should be repeated until the differences between consecutive iterations are specifically small and the estimators will converge to specific values. This iterative algorithm can be run in a statistical software such as S, S-Plus and R using the 'lme' function. The detailed theoretical discussion about this function has been presented in Pinheiro and Bates (2000).

2.3 Parameter Estimation using Area-level Data

For aggregated-level data, a similar function can be developed for parameter estimation based on the population model presented in equation (4). The area-level model for the

sample data is assumed to be derived by aggregating the unit-levels in the working model as follows:

$$\bar{y}_k = \bar{\mathbf{x}}_k'\beta + \epsilon_k \ ; \quad k = 1, \ldots, K \tag{21}$$

where:

$$\bar{\mathbf{x}}_k' = \begin{bmatrix} 1 & \bar{x}_{k1} & \bar{x}_{k2} & \ldots & \bar{x}_{kP} \end{bmatrix} \tag{22}$$

and $\epsilon_k = u_k + \bar{e}_k = \bar{y}_k - \bar{\mathbf{x}}_k'\beta$. Then, the log-likelihood function for the area-level model is given by:

$$l(\beta, \sigma_u^2, \sigma_e^2; \bar{\mathbf{y}}) = -\frac{1}{2}\left\{ ln(2K\pi) + ln\Big[det(\bar{\Sigma})\Big] + \epsilon'\bar{\Sigma}^{-1}\epsilon \right\} \tag{23}$$

where:

$$\epsilon' = [\epsilon_1 \ \epsilon_2 \ \ldots \ \epsilon_K] \qquad \& \qquad \bar{\Sigma} = diag\left(\sigma_u^2 + \frac{\sigma_e^2}{n_1}, \ \ldots \ , \sigma_u^2 + \frac{\sigma_e^2}{n_K}\right) . \tag{24}$$

Assuming the variance components to be known in the area-level model, the ML estimator for parameter β based on area-level sample data is:

$$\hat{\beta}^A = (\bar{\mathbf{x}}'\bar{\Sigma}^{-1}\bar{\mathbf{x}})^{-1}\bar{\mathbf{x}}'\bar{\Sigma}^{-1}\bar{\mathbf{y}} \tag{25}$$

where:

$$\bar{\mathbf{y}}' = [\bar{y}_1 \ \bar{y}_2 \ \ldots \ \bar{y}_K] \qquad \& \qquad \bar{\mathbf{x}}' = [\bar{\mathbf{x}}_1' \ \bar{\mathbf{x}}_2' \ \ldots \ \bar{\mathbf{x}}_K'] \ . \tag{26}$$

Fay and Herriot (1979) applied an area-level linear regression with area random effects in the case of unequal variances for predicting the mean value per capita income (PCI) in small geographical areas. The variance of the the sampling error is typically assumed to account for the complex sampling error for kth area and is considered be known in the Fay-Herriot model. This strong assumption seems unrealistic in practice.

Using area-level data, expressions for the Fisher scoring algorithm for the parameter λ is the same as in (20) (Longford, 2005; p.198). The initial value for σ_e^2 can be obtained from the unweighted OLS method. Then, using the Fisher scoring algorithm for the variance ratio, new estimated random effects for kth area in iteration $(t+1)$ can be calculated via:

$$\hat{\sigma}_{u(t+1)}^2 = \hat{\lambda}_{(t+1)}\hat{\sigma}_{e(t)}^2 \ . \tag{27}$$

Using $\hat{\sigma}^2_{u(t+1)}$ and $\hat{\sigma}^2_{e(t)}$, new estimators for $\hat{\bar{\Sigma}}_{(t+1)}$ and $\hat{\beta}^A_{(t+1)}$ can be be obtained. Then, a new estimated value for σ^2_e can be calculated as follows:

$$\hat{\sigma}^2_{e(t+1)} = \frac{1}{K-P}\ \hat{\epsilon}'_{(t+1)}\ \widehat{\bar{\mathbf{W}}}^{-1}_{(t+1)}\ \hat{\epsilon}_{(t+1)} \tag{28}$$

where, $\hat{\epsilon}_{(t+1)} = \left(\bar{\mathbf{y}} - \bar{\mathbf{x}}\hat{\beta}^A_{(t+1)}\right)$ and:

$$\widehat{\bar{\mathbf{W}}}_{(t+1)} = diag(\hat{\lambda}_{(t+1)} + \frac{1}{n_1},\ \ldots\ ,\hat{\lambda}_{(t+1)} + \frac{1}{n_K})\ . \tag{29}$$

Note that, the algorithm for calculating parameter estimates using individual and aggregated level analysis are very similar. The main difference is applied in calculating $\hat{\sigma}^2_{e(t+1)}$ using $\widehat{\mathbf{W}}_{(t+1)}$ with individual-level data and $\widehat{\bar{\mathbf{W}}}_{(t+1)}$ with aggregated-level data.

3 Synthetic and Empirical Best Liner Unbiased Predictor

Knowing estimates for regression parameters, the kth area mean for the target variable can be estimated based on the fitted statistical working models through the synthetic technique as follows:

$$\begin{aligned} \widehat{\bar{Y}}^{SU}_k &= \bar{\mathbf{X}}'_k\hat{\beta}^U \quad or \\ \widehat{\bar{Y}}^{SA}_k &= \bar{\mathbf{X}}'_k\hat{\beta}^A\ . \end{aligned} \tag{30}$$

Here, $\widehat{\bar{Y}}^{SU}_k$ and $\widehat{\bar{Y}}^{SA}_k$ respectively denote the unit-level and area-level mean synthetic estimators for the target variable within the kth area and $\bar{X}_k$ is the vector which includes population means of auxiliary variables. The estimated value for the parameter vector β using the individual-level sample data is $\hat{\beta}^U$ and $\hat{\beta}^A$ is the estimated value using the aggregated-level sample data.

In the general definition for Linear Mixed Model (LMM) presented in (3), $\mathbf{u}$ and $\mathbf{e}$ are assumed to be distributed independently with mean zero and covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively.

$$Var\begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}\ ,\quad E(\mathbf{e}) = \mathbf{0}\quad \&\quad E(\mathbf{u}) = \mathbf{0}\ . \tag{31}$$

The mean vector and covariance matrix for the target variable $\mathbf{Y}$ are respectively:

$$E(\mathbf{Y}|\mathbf{u}) = \mu_{\mathbf{Y}} = \mathbf{X}\beta + \mathbf{u} \quad \& \quad \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} \ . \tag{32}$$

The Best Linear Unbiased Estimation (BLUE) of the fixed effects β and Best Linear Unbiased Prediction (BLUP) of the random effects $\mathbf{u}$ in the LMM have been defined by Henderson (1950; 1975) as solutions to the following simultaneous equations.

$$\begin{aligned} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\tilde{\beta} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\tilde{\mathbf{u}} &= \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\tilde{\beta} + (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\tilde{\mathbf{u}} &= \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{aligned} \tag{33}$$

Robinson (1991) defined the BLUP as the best *linear* function of the data which is unbiased. Note that, within the statistical literature, it is conventional to use "estimation" for fixed effects and "predictions" for random effects. The results of these estimation methods are the *best*, as they minimize the generalized mean square error within the class of linear unbiased estimators, and they are *unbiased* as the average value of the estimates is equal to the average value of the quantity being estimated (Morris, 2001). Considering the equations in (33), $\mathbf{V}^{-1}$ can be defined in order to simplify the calculations as follows:

$$\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}. \tag{34}$$

It follows that:

$$\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}. \tag{35}$$

The plug-in formulas for $\tilde{\beta}$ and $\tilde{\mathbf{u}}$ can be calculated as a result of solving the equations above. These formulas are: [Morris, 2001]

$$\begin{aligned} \tilde{\beta} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \\ \tilde{\mathbf{u}} &= \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\beta}) \ . \end{aligned} \tag{36}$$

The ML estimator for the parameter vector β presented in (12) is then the same as the BLUE for this model parameter.

Under the general definition of LMM, prediction of a linear combination of the fixed and random effects ($\theta = \mathbf{b}'\beta + \mathbf{l}'\mathbf{u}$) has been discussed by Henderson (1975), Prasad and Rao

(1990), and Datta and Lahiri (2000). In a special case, the mentioned linear combination is presented as: $\mu_{\bar{Y}_k} = \bar{\mathbf{X}}_k'\beta + u_k$ when $\mathbf{b} = \bar{\mathbf{X}}_k$ and $\mathbf{l}' = (\underbrace{0, 0, \ldots, 0, 1}_{k}, 0, ..., 0)$. Then, the BLUP for this combination is:

$$\widetilde{\mu}_{\bar{Y}_k} = \bar{\mathbf{X}}_k'\tilde{\beta} + \mathbf{l}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\beta}) \ . \tag{37}$$

In the the case of the unit-level mixed model presented in (3), we have $\mathbf{G} = \sigma_u^2\,\mathbf{I}_K$ & $\mathbf{R} = \sigma_e^2\,\mathbf{I}_N$. In such a case, the linear combination of the predictions for fixed and random effects as presented by Henderson (1975) is:

$$\begin{aligned} \widetilde{\mu}_{\bar{Y}_k} = \bar{\mathbf{X}}_k'\tilde{\beta} + \tilde{u}_k &= \bar{\mathbf{X}}_k'\tilde{\beta} + \gamma_k(\bar{Y}_k - \bar{\mathbf{X}}_k'\tilde{\beta}) \\ &= \gamma_k\ \bar{Y}_k + (1-\gamma_k)\bar{\mathbf{X}}_k'\tilde{\beta} \end{aligned} \tag{38}$$

where:

$$\gamma_k = \frac{\sigma_u^2}{\sigma_u^2 + \psi_k} \qquad \& \qquad \psi_k = Var(\bar{e}_k|\bar{Y}_k) \ . \tag{39}$$

Considering the target of inference at the area-level, Ghosh and Rao (1994) defined the BLUP under the general LMM based on available sample data. Considering $\mu_{\bar{Y}_k} = E(\bar{Y}_k|u_k)$, the equation (38) can be based on available sample data as follows:

$$\begin{aligned} \widetilde{\mu}_{\bar{Y}_k} = \bar{\mathbf{X}}_k'\tilde{\beta} + \tilde{u} &= \bar{\mathbf{X}}_k'\tilde{\beta} + \gamma_k\Big(\bar{y}_k - \bar{\mathbf{x}}_k'\tilde{\beta}\Big) \\ &= \gamma_k\Big[\bar{y}_k + (\bar{\mathbf{X}}_k' - \bar{\mathbf{x}}_k')\tilde{\beta}\Big] + (1-\gamma_k)\bar{\mathbf{X}}_k'\tilde{\beta} \ . \end{aligned} \tag{40}$$

To calculate the BLUP value in equation (40), variance components are assumed to be known. Replacing the estimated values for the variance components in equation (40), a two-stage estimator will be obtained. The resulting estimator is presented by Harville (1991) as an "empirical BLUP" or EBLUP. The model parameters β, σ_e^2 and σ_u^2 can be empirically estimated for both individual or aggregated level analysis by the Fisher scoring algorithm as a general method for finding ML or REML parameter estimates, as presented in section 2.3.

Considering a true working model to be fitted on available sample data, an approximation for the Mean Square Error (MSE) of EPLUPs under general LMM is: (Saei and

Chambers, 2003b)

$$MSE_\xi(\widehat{\bar{Y}}^{EBLUP}) = MSE_\xi\left(\widetilde{\bar{Y}}\right) \simeq \mathcal{G}_1(\sigma) + \mathcal{G}_2(\sigma) + \mathcal{G}_3(\sigma) \tag{41}$$

where:

$$\begin{aligned}
&\mathcal{G}_1(\sigma) = (1-\gamma_k)\sigma_u^2 \\
&\mathcal{G}_2(\sigma) = (\bar{\mathbf{X}}_k - \gamma_k\bar{\mathbf{x}}_k)'\big[MSE_\xi(\tilde{\beta})\big](\bar{\mathbf{X}}_k - \gamma_k\bar{\mathbf{x}}_k) \\
&\mathcal{G}_3(\sigma) = \left(\frac{\sigma_e^2}{n_k}\right)^2\left(\sigma_u^2 + \frac{\sigma_e^2}{n_k}\right)^{-3} + \left[Var_\xi(\hat{\sigma}_u^2) + \frac{\sigma_u^4}{\sigma_e^4}Var_\xi(\hat{\sigma}_e^2) - 2\frac{\sigma_u^2}{\sigma_e^2}Cov_\xi(\hat{\sigma}_u^2, \hat{\sigma}_e^2)\right]
\end{aligned} \tag{42}$$

in which:

$$\gamma_k = \frac{\sigma_u^2}{\sigma_u^2 + \dfrac{\sigma_e^2}{n_k}} \quad \& \quad \sigma = (\sigma_u, \sigma_e).$$

The subscript ξ denotes the MSE, expectation and variance under the assumed population model. Considering model presented in (3) to be the actual population model, MSE of the resulting parameter estimate for β is as follows:

$$MSE_\xi\big(\tilde{\beta}\big) = Var_\xi\big(\tilde{\beta}\big) = (\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}$$

Replacing σ_u and σ_e respectively with $\widehat{\sigma}_u$ and $\widehat{\sigma}_e$, an estimation can be calculated for the equations presented in (42) as below:

$$\widehat{MSE}_\xi(\widehat{\bar{Y}}_k^{EBLUP}) = \widehat{MSE}_\xi\left(\widetilde{\bar{Y}}_k\right) \simeq \widehat{\mathcal{G}}_1(\sigma) + \widehat{\mathcal{G}}_2(\sigma) + 2\widehat{\mathcal{G}}_3(\sigma) \tag{43}$$

where:

$$\begin{aligned}
&\widehat{\mathcal{G}}_1(\sigma) = (1-\hat{\gamma}_k)\hat{\sigma}_u^2 \\
&\widehat{\mathcal{G}}_2(\sigma) = (\bar{\mathbf{X}}_k - \hat{\gamma}_k\bar{\mathbf{x}}_k)'\big[\widehat{MSE}_\xi(\tilde{\beta})\big](\bar{\mathbf{X}}_k - \hat{\gamma}_k\bar{\mathbf{x}}_k) \\
&\widehat{\mathcal{G}}_3(\sigma) = \left(\frac{\hat{\sigma}_e^2}{n_k}\right)^2\left(\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_k}\right)^{-3} + \left[\widehat{Var}_\xi(\hat{\sigma}_u^2) + \frac{\hat{\sigma}_u^4}{\hat{\sigma}_e^4}\widehat{Var}_\xi(\hat{\sigma}_e^2) - 2\frac{\hat{\sigma}_u^2}{\hat{\sigma}_e^2}\widehat{Cov}_\xi(\hat{\sigma}_u^2, \hat{\sigma}_e^2)\right]
\end{aligned} \tag{44}$$

in which:

$$\hat{\gamma}_k = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \dfrac{\hat{\sigma}_e^2}{n_k}} \ .$$

The additional term in (43) is due to: [Rao (2003), p. 104]

$$E\big[\mathcal{G}_3(\sigma)\big] = \mathcal{G}_1(\sigma) - \mathcal{G}_3(\sigma). \tag{45}$$

Detailed discussion about MSE of EBLUPs is presented by Prasad & Rao (1990) and Saei & Chambers (2003a).

4 Contextual model

It is well known that regression coefficients obtained from individual-level analysis can be different from those based on analysis of aggregate data. This is referred to as the ecological fallacy and can happen when the population model should include both unit-level and area-level fixed effects. It is common to derive mixed models at the individual level, but sometimes some area-level covariates may need to be included in the model.

In a contextual model, both individual level and group area-level covariates are included simultaneously (Mason *et al.* 1983 , 1984). The area-level covariates are referred to as 'contextual effects' and the model including both unit and area level covariates is a 'contextual model'. For example, the mean value of the auxiliary variable can be included in the statistical population model as the contextual effect as in:

$$\begin{aligned} Y_{ik} = \mathbf{X}_{ik}^{*'}\beta^* + u_k^* + e_{ik}^* \ ; \quad i = 1, \ldots, N_k \ \ \& \ \ k = 1, \ldots, K \\ u_k^* \overset{iid}{\sim} N(0, \sigma_{u^*}^2) \ \ ; \ \ e_{ik}^* \overset{iid}{\sim} N(0, \sigma_{e^*}^2). \end{aligned} \tag{46}$$

The aggregated form of this model is given as:

$$\begin{aligned} \bar{Y}_k = \bar{\mathbf{X}}_k' \beta^{**} + u_k^* + \bar{e}_k^* \\ u_k^* \overset{iid}{\sim} N(0, \sigma_{u^*}^2) \ \ ; \ \ \bar{e}_k^* = \frac{1}{N_k} \sum_{i=1}^{N_k} e_{ik}^* \sim N(0, \frac{\sigma_{e^*}^2}{N_k}). \end{aligned} \tag{47}$$

Here,

$$\mathbf{X}_{ik}^{*'} = [\mathbf{X}_{ik}' \mid \bar{\mathbf{X}}_k'],$$
$$\left(\beta^{*I}\right)' = [\beta_0^{*I} \quad \beta_1^{*I} \quad \ldots \quad \beta_P^{*I}] \quad , \quad \left(\beta^{*c}\right)' = [\beta_0^{*C} \quad \beta_1^{*C} \quad \ldots \quad \beta_P^{*C}]$$

$$\beta^{*'} = \left[\left(\beta^{*I}\right)' \mid \left(\beta^{*C}\right)'\right] \quad \& \quad \beta^{**} = \begin{bmatrix} \beta_0^{*I} + \beta_0^{*C} \\ \beta_1^{*I} + \beta_1^{*C} \\ \beta_2^{*I} + \beta_2^{*C} \\ \vdots \\ \beta_P^{*I} + \beta_P^{*C} \end{bmatrix} . \tag{48}$$

Contextual models help researchers to understand and study the issue of the ecological fallacy which occurs when researchers want to draw a conclusion about an individual-level relationship based on aggregated-level data analysis. This causes an error in the interpretation of statistical data as the results based on purely aggregated-level analysis may not be appropriate for inference about an individual based characteristic (Seiler and Alvarez, 2000). When contextual effects exist in the population model but are ignored in working models, the resulting regression coefficient estimates from unit-level and area-level sample data will be different in expectation. This is referred to as an ecological fallacy.

When area means appear in the population model as contextual effects, the resulting correct model for the sample unit-level data is:

$$\begin{gathered} y_{ik} = \mathbf{X}_{(s)ik}^{*'}\beta^* + u_k^* + e_{ik}^* \ ; \quad i = 1, \ldots, N_k \quad \& \quad k = 1, \ldots, K \\ u_k^* \overset{iid}{\sim} N(0, \sigma_{u^*}^2) \quad ; \quad e_{ik}^* \overset{iid}{\sim} N(0, \sigma_{e^*}^2) \end{gathered} \tag{49}$$

and the true model for aggregate sample data is:

$$\begin{gathered} \bar{y}_k = \bar{\mathbf{X}}_{(s)k}^{*'}\beta^{**} + u_k^* + \bar{e}_k^* \\ u_k^* \overset{iid}{\sim} N(0, \sigma_{u^*}^2) \quad ; \quad \bar{e}_k^* = \frac{1}{n_k}\sum_{i=1}^{n_k} e_{ik}^* \sim N(0, \frac{\sigma_{e^*}^2}{n_k}) \end{gathered} \tag{50}$$

where:

$$\begin{gathered} \mathbf{X}_{(s)ik}^{*'} = [\mathbf{x}_{ik}' \mid \bar{\mathbf{X}}_k'] \\ \bar{\mathbf{X}}_{(s)k}^{*'} = [\bar{\mathbf{x}}_k' \mid \bar{\mathbf{X}}_k']. \end{gathered} \tag{51}$$

In the next section, we will consider how small area estimates based on unit-level working model (5) and the aggregate working model (21) perform when the area means are included in the assumed population model as contextual effects.

5 Numerical Simulation Study

This section presents the results of a simulation study to assess the empirical MSE of synthetic estimators and EBLUPs based on unit-level and area-level mixed models. The population data in this study has been generated based on available area information available in Australia. There are six states and two mainland territories in Australia and each has been divided, thereby forming a total of 57 statistical sub-divisions.

As a hypothetical example, we suppose that there is interest in the mean value of income for the 57 statistical sub-divisions within Australia. It is assumed that there is a linear relationship between the weekly gross salary as the variable of interest and the weekly hours worked for individuals aged 15 and over. In this simulation, population data is generated based on the contextual model (46) using parameter values obtained on the relation between weekly gross salary and hours worked for individuals over 15 in the Australian census 2006. Sample units are then selected from different areas based on a stratified random sampling design in which the sample sizes in the 57 areas are allocated proportionally to their population sizes. Table (1) presents the model parameter values used in generating the population of individuals.

Table 1: Parameter Values Considered in the Population Model

$\beta' = \begin{bmatrix}\beta_0^* & \beta_1^{*I} & \beta_1^{*C}\end{bmatrix}$	σ_u^*	σ_e^*	λ^*
[-123.61 14.93 3.77]	114.3530	384.6394	0.884

A total population size of 16278397 individuals was generated corresponding to the population aged over 15 in the Australian 2006 Census. A total sample of size $n = 2133$ was then selected based on the determined design. The sampling process was repeated 1000 times in this study. For each sample synthetic estimators and EBLUPs are then

estimated based on the two working models presented in Table 2, corresponding to models presented in (5) and (21). Details of the population and sample sizes are given in Table 4 and 5.

Table 2: Summery of Working Models and Predictors

Working Models	Synthetic Estimator	EBLUP
$y_{ik}^{(W_1)} = \mathbf{x}'_{ik}\beta + u_k + e_{ik}$	$\bar{\mathbf{X}}'_k\hat{\beta}$	$\bar{\mathbf{X}}'_k\hat{\beta} + \hat{u}_k$
$\bar{y}_k^{(W_2)} = \bar{\mathbf{x}}'_k\beta^{**} + u_k^* + \bar{e}_k^*$	$\bar{\mathbf{X}}'_k\hat{\beta}^{**}$	$\bar{\mathbf{X}}'_k\hat{\beta}^{**} + \hat{u}_k^*$

Assuming the contextual model presented in (46) applies for the population, fitting working model W_1 leads to biased parameter estimates. For the aggregate data the true sampling model is (50), parameter estimates based on W_2 may also be biased as sample area means ($\bar{\mathbf{x}}_k$) and population area means ($\bar{\mathbf{X}}_k$) may differ. However, W_2 includes $P+1$ regression coefficients to be estimated while $2P+1$ regression coefficients are included in models (49) and (50). Therefore, the dimension reduction in calculating model parameter estimates is an advantage of applying W_2.

Synthetic estimates and EBLUPs of the area means are then calculated based on W_1 and W_2 being fitted on the sample data. This allows a comparison to be made among unit-level and area-level working models introduced in Table 2 when area means are involved in the population model as a contextual effect.

The target of inference is $\bar{Y}_k$ given by (47). The bias of the unit-level synthetic estimate is:

$$Bias_\xi\left(\widehat{\bar{Y}}_k^{SU}\right) = E_\xi(\widehat{\bar{Y}}_k^{Syn(W_1)} - \bar{Y}_k) = \bar{\mathbf{X}}'_k E_\xi[\hat{\beta}^U - \beta^{**}] \ , \tag{52}$$

and for the area-level synthetic estimator the bias is:

$$Bias_\xi\left(\widehat{\bar{Y}}_k^{SA}\right) = E_\xi(\widehat{\bar{Y}}_k^{Syn(W_2)} - \bar{Y}_k) = \bar{\mathbf{X}}'_k E_\xi[\hat{\beta}^A - \beta^{**}] \ . \tag{53}$$

It can be shown that $E_\xi(\hat{\beta}^U) \approx \beta$ and $E_\xi(\hat{\beta}^U - \beta^{**}) \approx [0 \quad \beta^{*C}]'$. Therefore, the bias of the unit-level synthetic estimator for kth area is $\bar{\mathbf{X}}_k\beta^*$. For $\hat{\beta}^A$, the components of β^{**} associated with β^{*I} are unbiasedly estimated and the components associated with β^{*C} are subject to attenuation because of the difference between $\bar{\mathbf{x}}$ and $\bar{\mathbf{X}}$. However, we would

expect the attenuation not to completely eliminate the component associated with β and therefore $\hat{\beta}^A$ to be a less biased estimate of β^{**} than $\hat{\beta}^U$.

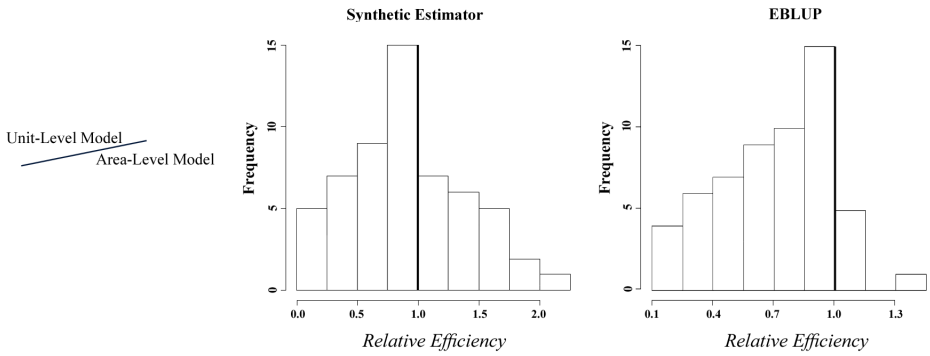
The bias of the unit-level EBLUP for kth area mean is calculated as follows:

$$Bias_\xi\left(\widetilde{\bar{Y}}_k^{(W_1)}\right) = \left[\bar{\mathbf{X}}_k' - E_\xi(\widehat{\gamma}_k)\bar{\mathbf{x}}_k'\right] E_\xi\left(\tilde{\beta}^{(W_1)} - \beta^{**}\right) + Cov_\xi\left[\widehat{\gamma}_k, \left(\bar{y}_k - \bar{\mathbf{x}}_k'\tilde{\beta}^{(W_1)}\right)\right]. \quad (54)$$

We see that the first term reduces the bias compared with the unit-level synthetic estimation. The second term should be negligible. A similar result holds for area-specific EBLUP obtained from the appropriate aggregate working model, W_2.

Figure 1 summarizes the empirical results by giving the ratio of MSEs for the SAEs based on unit-level and area-level models for the *57* areas in the simulation. When a contextual effect is present in the assumed population model, the ratio varies below and above *1* for the synthetic method, but is generally below *1* for the resulting EBLUPs. The variance of estimators obtained based on the individual-level analysis are less than the variance in the aggregated-level approach. However, the resulting bias in the estimation of β^{**} is greater. Using the synthetic method in this simulation, for about half the areas the area-level approach is better than the unit-level approach in terms of MSE. However, when the EBLUP is applied, the reduction in biases leads to the unit-level approach having lower MSE in all but a few areas.

Figure 1: The Relative Efficiency of Unit-level Model to Area-level Model

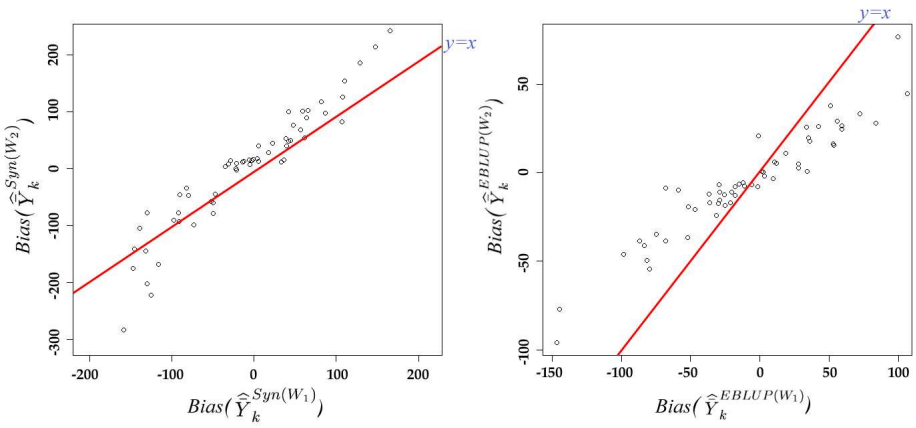


A comparison between the resulting bias based on the synthetic estimation approach and EBLUP technique is presented in Figure 2 for the target areas. For positive biases of the synthetic estimates, unit-level and area-level results look similar in terms of bias values. However, when the resulting biases for unit-level synthetic estimates are negative, less biased synthetic estimates can be calculated based on area-level models. For calculated EBLUPs the bias of the unit-level estimates are predominately larger than that of aggregated-level estimates. The bias seemes to be decreased in unit-level estimation based on the EBLUP technique comparing with the synthetic estimation method. This is due to reduced weight given to the regression component in the presented EBLUP technique. Ignoring the difference between the sample and population area means for the auxiliary variable in kth area, the bias for unit-level synthetic estimator and EBLUP for kth area mean are

$$\begin{aligned} Bias_{\xi}\left(\tilde{\bar{Y}}_k^{(W_1)}\right) &\approx (1-\gamma_k)\bar{\mathbf{X}}_k' Bias_{\xi}\left(\tilde{\beta}^{(W_1)}\right) = \left(\frac{\frac{\sigma_e^2}{n_k}}{\sigma_u^2+\frac{\sigma_e^2}{n_k}}\right)\bar{\mathbf{X}}_k' Bias_{\xi}\left(\tilde{\beta}^{(W_1)}\right) \\ Bias_{\xi}\left(\tilde{\bar{Y}}_k^{(SU)}\right) &\approx \bar{\mathbf{X}}_k' Bias_{\xi}\left(\tilde{\beta}^{(W_1)}\right). \end{aligned} \tag{55}$$

As shown in (55), there is less bias in the unit-level EBLUP comparing with the unit-level synthetic estimator for kth area. This reduction depends on n_k.

Figure 2: Resulting Bias for Synthetic Estimators and EBLUPs



Means and variances of the parameter estimates for working models used in this numerical study are presented in Table 3. As expected, estimated values for the intercept and slope are less biased in the aggregated-level analysis. We see that the unit-level slope estimate is unbiased for β_1, and the area-level slope estimate is closer to $\beta_1^{*I} + \beta_1^{*C} = \beta_1^{**}$, but still smaller, consistent with the attenuation effect noted above. As expected, the standard error of all the parameter estimates are larger for area-level analysis. Interestingly, the bias for the estimate of λ appears to be less for the area-level approach. The generally smaller bias of the area-level analysis but larger MSEs, suggests that existing contextual effects in the population model being considered in W_2 causes less bias of parameter estimates with smaller bias comparing with that of W_2.

Table 3: Parameter Estimates under Population 2

	W_1		W_2
$\bar{\hat{\beta}}$	$\begin{pmatrix} 71.14 \\ 13.78 \end{pmatrix}$	$\bar{\hat{\beta}}^{**}$	$\begin{pmatrix} -88.71 \\ 17.29 \end{pmatrix}$
$Bias(\bar{\hat{\beta}})$	$\begin{pmatrix} 18.74 \\ -4.92 \end{pmatrix}$	$Bias(\bar{\hat{\beta}}^{**})$	$\begin{pmatrix} 11.10 \\ -2.07 \end{pmatrix}$
$SE(\bar{\hat{\beta}})$	$\begin{pmatrix} 7.83 \\ 0.71 \end{pmatrix}$	$SE(\bar{\hat{\beta}}^{**})$	$\begin{pmatrix} 11.94 \\ 4.02 \end{pmatrix}$
$\bar{\hat{\sigma}}_u$	129.45	$\bar{\hat{\sigma}}_u^*$	51.47
$Bias(\bar{\hat{\sigma}}_u)$	7.99	$Bias(\bar{\hat{\sigma}}_u^*)$	-17.47
$SE(\bar{\hat{\sigma}}_u)$	6.18	$SE(\bar{\hat{\sigma}}_e^*)$	21.41
$\bar{\hat{\sigma}}_e$	285.36	$\bar{\hat{\sigma}}_e^*$	369.07
$Bias(\bar{\hat{\sigma}}_e)$	-26.72	$Bias(\bar{\hat{\sigma}}_e^*)$	-7.49
$SE(\bar{\hat{\sigma}}_e)$	17.50	$SE(\bar{\hat{\sigma}}_e^*)$	24.08
$\bar{\lambda}$	0.112	$\bar{\lambda}^*$	0.074
$Bias(\bar{\lambda})$	0.010	$Bias(\bar{\lambda}^*)$	0.007
$SE(\bar{\lambda})$	0.022	$SE(\bar{\lambda}^*)$	0.071

6 Conclusion

The goal of this paper is to evaluate SAE techniques based on statistical models at different levels and to study the effect of possible area-level contextual effects in the population model. The possible effects of ignoring these important area-level factors is explained for unit-level working models being fitted on sample data. In order to consider realistic situations, individual-level data from the Australian 2006 Census are used to estimate the parameter values in population model.

If unit-level data are available, information from individuals can be used in the working model. Estimators can then be obtained at the area level using aggregating techniques. If data are unaccessible for unit-level modeling while area-level data are available, area-level models can be developed for aggregate-level analysis and parameters used in producing estimates at district levels are estimated from an area-level model directly. When area means appear in the unit-level population model as contextual effects but are ignored in

the individual-level working model, the resulting parameter estimates are biased while the area-level model will automatically include these effects in estimation. In such a case, the resulting parameter estimates would be unbiased or less biased, and an area-level analysis may be preferable even if individual-level data are available.

Choosing individual-level analysis helps to produce small area estimates with smaller variances. However, if the unit-level model is misspecified by exclusion of important auxiliary variables, parameter estimates obtained from the individual and aggregate-level analysis will have different expectations. In particular, if an important contextual variable is omitted, the parameter estimates obtained from an individual-level analysis will be biased, whereas an aggregated-level analysis can produce less biased estimates. Even if contextual variables are included in an individual-level analysis, there may be an increase in the variance of parameter estimates due to the increased number of variables in the population model.

The size of the contextual effect will be an important feature in determining the relative efficiency of unit-level and area-level approaches. When individual-level analysis is being used, the theory and empirical results suggest using EBLUP technique as it is more efficient than the synthetic method.

References

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

Datta, G. S., and Lahiri, P. (2000). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems. *Statistica Sinica.* **10**, 613-627.

Fay, R. E., and Herriot, R. A. (1979). Estimates of Income for Small Places: an Application of James-Stein Procedures to Census Data. *Journal of The American Statistical Association.* **74**, 269-277.

Harville, D. A. (1991). That BLUP is a Good Thing: The Estimation of Random Effects, (Comment). *Statistical Science.* **6**, 35-39.

Henderson, C. R., (1950). Estimation of Genetic Parameters (abstract). *The Annals of*

Mathematical Statistics. **21**, 309-310.

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics.* **31**, 423-447.

Ghosh, M., and Rao, J. N. K. (1994) Small Area Estimation: an Appraisal. *Statistical Science.* **9**, 55-93.

Khoshgooyanfard, A., and Taheri Monazah, M. (2006). A Cost-Effective Strategy for Provincial Unemployment Estimation: A Small Area Approach. *Survey Methodology.* **32**, 105-114.

Longford, N. T. (2005). *Missing Data and Small Area Estimation.* Spring-Verlag.

Longford, N. T. (1993). *Random coefficient models.* Oxford University Press; New York.

Mason, W. M., Wong, G. Y., and Entwisle, B. (1983 - 1984). Contextual Analysis through the Multilevel Linear Model. *Sociological Methodology.* **14**. 72-103.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*, 2nd Edition. Wiley.

Morris, J. S. (2002). The BLUPs are not best when it comes to bootstrapping. Statistics and Probability Letters. **56**. 425-430.

Osborne, M. R. (1992). Fisher's method of scoring. *International Statistical Review.* **60**, 99-117.

Pinheiro, J. C., and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS.* Springer; New York.

Prasad, N. G. N., and Rao, J. N. K. (1990). The Estimation of Mean Squared Errors of Small Area Estimators. *Journal of the American Statistical Association.* **85**, 163-171.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley; New York.

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science.* **6**, 15-32.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression* . Cambridge University Press.

Saei, A., and Chambers, R. (2003a). 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.

Saei, A., and Chambers, R. (2003b). Small area estimation: A review of methods based on the application of mixed models. *S^3RI Methodology Working Paper M03/16.*; University of Southampton.

Seiler, F. A., and Alvarez, J. L. (2000). Is the Ecological Fallacy a Fallacy? *Human and Ecological Risk Assessment.* **6**, 921-941.

Table 4: The Population Size for Different Statistical Subdivisions

STATE	No.	Statistical Subdivisions	Population(15 an over)	Total
ACT	1	Canberra	276469	**276469**
NSW	2	Murray	141384	**5554876**
	3	Northern	207344	
	4	Murrumbidgee	179500	
	5	Sydney	2643880	
	6	Richmond-Tweed	301849	
	7	South Eastern	211561	
	8	Central West	123473	
	9	Mid-North Coast	351211	
	10	Illawarra	541424	
	11	Hunter	707457	
	12	Far West	26961	
	13	North Western	118832	
NT	14	Northern Territory - Bal	74040	**163164**
	15	Darwin	89124	
QLD	16	Brisbane	1481729	**2942559**
	17	Central West	7683	
	18	Far North	189129	
	19	South West	13461	
	20	Fitzroy	112659	
	21	Moreton	427387	
	22	North West	20137	
	23	Mackay	125319	
	24	Wide Bay-Burnett	226345	
	25	Northern	159776	
	26	Darling Downs	178934	
SA	27	Adelaide	947857	**1244878**
	28	Outer Adelaide	93348	
	29	Northern	65062	
	30	Murray Lands	55298	
	31	Eyre	28617	
	32	Yorke and Lower North	37557	
	33	South East	17139	
TAS	34	Northern	112182	**390217**
	35	Greater Hobart	166825	
	36	Mersey-Lyell	81914	
	37	Southern	29296	
VIC	38	Melbourne	3038339	**4138085**
	39	Central Highlands	121149	
	40	Ovens-Murray	78547	
	41	Gippsland	135565	
	42	Goulburn	159950	
	43	Mallee	75144	
	44	Loddon	143693	
	45	Barwon	221846	
	46	Wimmera	37877	
	47	Western District	57861	
	48	East Gippsland	68114	
WA	49	Lower Great Southern	41606	**1568149**
	50	Perth	1246870	
	51	Pilbara	11127	
	52	South West	111080	
	53	South Eastern	45401	
	54	Upper Great Southern	13544	
	55	Central	31724	
	56	Kimberley	26603	
	57	Midlands	40194	

Table 5: The Sample Size for Different Statistical Subdivisions

STATE	No.	Statistical Subdivisions	Sample Size	Total
ACT	1	Canberra	36	**36**
NSW	2	Murray	19	**730**
	3	Northern	27	
	4	Murrumbidgee	23	
	5	Sydney	347	
	6	Richmond-Tweed	40	
	7	South Eastern	28	
	8	Central West	16	
	9	Mid-North Coast	46	
	10	Illawarra	71	
	11	Hunter	93	
	12	Far West	4	
	13	North Western	16	
NT	14	Northern Territory - Bal	10	**22**
	15	Darwin	12	
QLD	16	Brisbane	194	**386**
	17	Central West	1	
	18	Far North	25	
	19	South West	2	
	20	Fitzroy	15	
	21	Moreton	56	
	22	North West	3	
	23	Mackay	16	
	24	Wide Bay-Burnett	30	
	25	Northern	21	
	26	Darling Downs	23	
SA	27	Adelaide	121	**160**
	28	Outer Adelaide	12	
	29	Northern	9	
	30	Murray Lands	7	
	31	Eyre	4	
	32	Yorke and Lower North	5	
	33	South East	2	
TAS	34	Northern	15	**52**
	35	Greater Hobart	22	
	36	Mersey-Lyell	11	
	37	Southern	4	
VIC	38	Melbourne	398	**542**
	39	Central Highlands	16	
	40	Ovens-Murray	10	
	41	Gippsland	18	
	42	Goulburn	21	
	43	Mallee	10	
	44	Loddon	18	
	45	Barwon	29	
	46	Wimmera	5	
	47	Western District	8	
	48	East Gippsland	9	
WA	49	Lower Great Southern	5	**205**
	50	Perth	163	
	51	Pilbara	1	
	52	South West	15	
	53	South Eastern	6	
	54	Upper Great Southern	2	
	55	Central	4	
	56	Kimberley	4	
	57	Midlands	5	