2013

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, dsteel@uow.edu.au

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

01-13

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

Mohammad-Reza Namazi-Rad and David Steel

National Institute for Applied Statistics Research Australia, 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*§, and David Steel*

* *Institute for Applied Statistics Research Australia, University of Wollongong, NSW 2522, Australia*

§ *SMART Infrastructure Facility, University of Wollongong, NSW 2522, Australia*

January 2013

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. 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 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 using 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. 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), Longford (2005), Lehtonen & Veijanen (2009), and Datta (2009) 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 mixed 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 use of unit-level model includes area-level inference, as well. 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 of covariates, 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 motivated by actual census data, 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 relate the variable of interest to a set of covariates for which data is collected in the survey and auxiliary population information is available for each target sample area. Parameters of the model can then be estimated using data for the entire sample which can be combined with the auxiliary information available for each small area to produce small area estimates. Efficient models usually include random effects to explain the variation between target areas within the population that is not explained by the covariates available (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 linear mixed model for the population 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 the 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. The area-level linear mixed model obtained from (1) for the population area means is given as:

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

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:
Unit-Level Population Model :

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$
$$\mathbf{u} \sim N(\mathbf{0},\; \sigma_u^2 \; \mathbf{I}_K) \quad ; \quad \mathbf{e} \sim N(\mathbf{0},\; \sigma_e^2 \; \mathbf{I}_N) \tag{3}$$

Area-Level Population Model :

$$\bar{\mathbf{Y}} = \bar{\mathbf{X}}\beta + \mathbf{u} + \bar{\mathbf{e}}$$
$$\mathbf{u} \sim N(\mathbf{0},\; \sigma_u^2 \; \mathbf{I}_K) \quad ; \quad \bar{\mathbf{e}} \sim N\big(\mathbf{0},\; diag(\tfrac{\sigma_e^2}{N_1}, \;\; \ldots, \;\; \tfrac{\sigma_e^2}{N_K})\big). \tag{4}$$

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 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, a sample s of size n is assumed to be selected from the target population U. The part of the overall sample s which falls into the kth area is $s_k = s \bigcap U_k$ and is of size n_k.

A direct estimate for a target small area is based only on the available data for that area. 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:

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

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 $\mathbf{x}_k$. Here, $\mathbf{z}$ 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. We assume that the sampling scheme used is noninformative, so the same model can be used for the sample and population at the individual level. We have assumed that there is at least one sample member in each small area, although the situation where some small areas have no sample units is easily handled.

For the model given by (5), the likelihood is:

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

where c is a constant and Σ is the block-diagonal variance-covariance matrix given as: $\Sigma = diag(\Sigma_k)$ where: $\Sigma_k = \sigma_u^2 \mathbf{J}_{n_k} + \sigma_e^2 \mathbf{I}_{n_k}$ & $\mathbf{J}_{n_k} = \mathbf{1}_{n_k} \mathbf{1}'_{n_k}$. Let $l(\beta, \sigma_u^2, \sigma_e^2;\ \mathbf{y})$ to be the associated log-likelihood function:

$$l(\beta, \sigma_u^2, \sigma_e^2;\ \mathbf{y}) = ln(c) - \frac{1}{2}\sum_{k=1}^{K} ln|\Sigma_k| - \frac{1}{2}\sum_{k=1}^{K} \varsigma_k'\ \Sigma_k^{-1} \varsigma_k \tag{7}$$

where:

$$\varsigma_k = \mathbf{y} - \mathbf{x}\beta \quad \& \quad \Sigma_k^{-1} = \sigma_e^{-2}(\mathbf{I}_{n_k} - \tfrac{\gamma_k}{n_k}\ \mathbf{1}_{n_k}\ \mathbf{1}'_{n_k}) \tag{8}$$

in which:

$$\gamma_k = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{n_k}}\ . \tag{9}$$

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:

$$\hat{\beta}^U = (\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}\mathbf{x}'\Sigma^{-1}\mathbf{y} \tag{10}$$

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

Longford (1993) considers the Fisher scoring algorithm for estimating a value for parameter θ:

$$\theta_{(t+1)} = \theta_{(t)} + \mathcal{I}^{-1}(\theta_{(t)})\ \mathcal{S}(\theta_{(t)}) \tag{11}$$

where:

$$\mathcal{I}(\theta_{(t)}) = -E\Big(\frac{\partial^2 l}{\partial\theta\ \partial\theta'}\Big)\Big|_{\theta=\theta(t)} \quad \& \quad \mathcal{S}(\theta_{(t)}) = \frac{\partial l}{\partial \theta}\Big|_{\theta=\theta(t)} \tag{12}$$

The notations (t) and $(t{+}1)$ denote the previous and new estimated values for these parameters, respectively. Longford (1993) suggests a reparametrization using the variance ratio $\lambda = \sigma_u^2/\sigma_e^2$, so $\theta^* = (\beta, \sigma_e^2, \lambda)$. For the parameter λ,

$$\frac{\partial l(\theta^*;\ \mathbf{y})}{\partial \lambda} = -\frac{1}{2}\sum_{k=1}^{K} \mathbf{1}'_{n_k}\mathbf{W}_k^{-1}\mathbf{1}_{n_k} + \frac{1}{2\sigma_e^2}\sum_{k=1}^{K}\Big(\varsigma_k'\mathbf{W}_k^{-1}\mathbf{1}_{n_k}\Big)^2 \tag{13}$$

and,

$$-E\Big(\frac{\partial^2 l(\theta^*;\ \mathbf{y})}{\partial^2 \lambda}\Big) = \frac{1}{2}\sum_{k=1}^{K}\big(\mathbf{1}'_{n_k}\mathbf{W}_k^{-1}\mathbf{1}_{n_k}\big)^2 = \frac{1}{2}\sum_{k=1}^{K}\Big(f_k^{-1}\mathbf{1}'_{n_k}\mathbf{1}_{n_k}\Big)^2$$

$$-E\Big(\frac{\partial^2 l(\theta^*;\ \mathbf{y})}{\partial\beta\ \partial\lambda}\Big) = \mathbf{x}'\frac{\partial \mathbf{W}^{-1}}{\partial\lambda}\ E(e_{ik}) = 0 \tag{14}$$

where $f_k = 1 + n_k\lambda$ and

$$\mathbf{W} = \sigma_e^{-2}\Sigma \quad ; \quad \mathbf{W}_k = \sigma_e^{-2}\big(\sigma_u^2\ \mathbf{1}_{n_k}\ \mathbf{1}'_{n_k} + \sigma_e^2\ \mathbf{I}_{n_k}\big) = \lambda\ \mathbf{1}_{n_k}\mathbf{1}'_{n_k} + \mathbf{I}_{n_k}$$

$$\mathbf{W}^{-1} = \sigma_e^2\ \Sigma^{-1} \quad ; \quad \mathbf{W}_k^{-1} = \frac{-\sigma_u^2}{\sigma_e^2 + n_k\sigma_u^2}\ \mathbf{1}_{n_k}\ \mathbf{1}'_{n_k} + \mathbf{I}_{n_k}\ . \tag{15}$$

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

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

$$= \hat{\lambda}_{(t)} + \Big[\frac{1}{2}\sum_{k=1}^{K}\frac{n_k^2}{f_{k(t)}^2}\Big]^{-1}\Big[-\frac{1}{2}\sum_{k=1}^{K}(\frac{n_k}{f_{k(t)}}) + \frac{1}{2\hat{\sigma}^2_{e(t)}}\sum_{k=1}^{K}(f_{k(t)}^{-1}\ \hat{\varsigma}'_{k(t)}\ \mathbf{1}_{n_k})^2\Big] \tag{16}$$

where $f_{k(t)} = 1 + n_k\lambda_{(t)}$, and $\hat{\varsigma}_{k(t)} = \mathbf{y}_k - \mathbf{x}'_k\hat{\beta}^U_{(t)}$. Initial values can be based on ordinary least squares estimates. For the other parameters in θ^*,

$$\begin{aligned}\hat{\beta}_{(t+1)} &= (\mathbf{x}'\hat{\Sigma}^{-1}_{(t+1)}\ \mathbf{x})^{-1}\mathbf{x}'\hat{\Sigma}^{-1}_{(t+1)}\ \mathbf{y}\\ \hat{\sigma}^2_{e(t+1)} &= \hat{\varsigma}'_{(t+1)}\widehat{W}^{-1}_{(t+1)}\hat{\varsigma}_{(t+1)},\end{aligned} \tag{17}$$

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

Given the estimates of β and σ_e^2, the sample data only affect the calculation in equation (16) 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. Detailed discussion on this estimation approach is presented by Pinheiro and Bates (2000).

2.3 Parameter Estimation using Area-level Data

For aggregated-level data, a similar approach can be developed for parameter estimation. The area-level model for the sample data is assumed to be derived by aggregating the unit-level working model given by (5) as follows:

$$\bar{y}_k = \bar{\mathbf{x}}'_k \beta + \epsilon_k \tag{18}$$

where:

$$\bar{\mathbf{x}}'_k = \begin{bmatrix} 1 & \bar{x}_{k1} & \bar{x}_{k2} & \dots & \bar{x}_{kP} \end{bmatrix} \tag{19}$$

and $\epsilon_k = u_k + \bar{e}_k$. In the matrix form the model is:

$$\bar{\mathbf{y}} = \bar{\mathbf{x}}'\beta + \epsilon \tag{20}$$

where,

$$\bar{\mathbf{y}}' = [\bar{y}_1 \;\; \bar{y}_2 \;\; \dots \;\; \bar{y}_K] \;\; , \;\; \bar{\mathbf{x}}' = [\bar{\mathbf{x}}_1 \;\; \bar{\mathbf{x}}_2 \;\; \dots \;\; \bar{\mathbf{x}}_K] \;\; \& \;\; \epsilon' = [\epsilon_1 \; \epsilon_2 \; \dots \; \epsilon_K]. \tag{21}$$

Then, the log-likelihood function for the area-level model is given by:

$$l(\beta, \sigma_u^2, \sigma_e^2; \bar{\mathbf{y}}) = -\frac{1}{2}\Big\{ ln(2K\pi) + ln\Big[det(\bar{\Sigma})\Big] + \epsilon'\bar{\Sigma}^{-1}\epsilon \Big\} \tag{22}$$

where, $\bar{\Sigma} = diag\Big(\sigma_u^2 + \dfrac{\sigma_e^2}{n_1}, \; \dots \; , \sigma_u^2 + \dfrac{\sigma_e^2}{n_K} \Big)$.

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:

$$\hat{\beta}^A = (\bar{\mathbf{x}}'\bar{\Sigma}^{-1}\bar{\mathbf{x}})^{-1}\bar{\mathbf{x}}'\bar{\Sigma}^{-1}\bar{\mathbf{y}}. \tag{23}$$

Fay and Herriot (1979) applied an area-level linear regression to survey estimates 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 the survey estimates for the kth area and is considered be known in the Fay-Herriot model. However, 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 (16) (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:

$$\hat{\sigma}^2_{u(t+1)} = \hat{\lambda}_{(t+1)}\hat{\sigma}^2_{e(t)} \ . \tag{24}$$

Using $\hat{\sigma}^2_{u(t+1)}$ and $\hat{\sigma}^2_{e(t)}$, new estimators for $\hat{\bar{\Sigma}}_{(t+1)}$ and $\hat{\beta}^A_{(t+1)} = \left(\bar{\mathbf{x}}'\hat{\bar{\Sigma}}^{-1}_{(t+1)}\bar{\mathbf{x}}\right)^{-1}\bar{\mathbf{x}}'\hat{\bar{\Sigma}}^{-1}_{(t+1)}\bar{\mathbf{y}}$ can be be obtained. Then, a new estimated value for σ^2_e can be calculated as follows:

$$\hat{\sigma}^2_{e(t+1)} = \frac{1}{K-P}\ \hat{\epsilon}'_{(t+1)}\ \widehat{\bar{\mathbf{W}}}^{-1}_{(t+1)}\ \hat{\epsilon}_{(t+1)} \tag{25}$$

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

$$\widehat{\bar{\mathbf{W}}}_{(t+1)} = diag(\hat{\lambda}_{(t+1)} + \frac{1}{n_1},\ \ldots\ ,\hat{\lambda}_{(t+1)} + \frac{1}{n_K}) \ . \tag{26}$$

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

Given 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:

$$\widehat{\bar{Y}}^{SU}_k = \bar{\mathbf{X}}'_k\hat{\beta}^U \quad or \quad \widehat{\bar{Y}}^{SA}_k = \bar{\mathbf{X}}'_k\hat{\beta}^A \ . \tag{27}$$

Here, $\widehat{\bar{Y}}^{SU}_k$ and $\widehat{\bar{Y}}^{SA}_k$ respectively denote the unit-level and area-level synthetic estimators for the kth area mean and $\bar{\mathbf{X}}_k$ is the vector which includes population means of auxiliary variables.

For the Linear Mixed Model (LMM) presented in (3), the Best Linear Unbiased Estimation (BLUE) of the fixed effects β and Best Linear Unbiased Prediction (BLUP) of the random effects $\mathbf{u}$ have been defined by Henderson (1950; 1975) and Morris (2002) as follows:

$$\tilde{\beta} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y} \quad \& \quad \tilde{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\Sigma^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\beta}) \ , \tag{28}$$

where $\mathbf{G} = \sigma_u^2\, \mathbf{I}_K$. The ML estimator for the parameter vector β presented in (10) is then the same as the BLUE for this model parameter.

For the LMM, prediction of a linear combination of the fixed and random effects $\mathbf{b}'\beta + \mathbf{l}'\mathbf{u}$ has been discussed by Henderson (1975), Prasad and Rao (1990), and Datta and Lahiri (2000). For the special case $\mu_{\bar{Y}_k} = \bar{\mathbf{X}}_k'\beta + u_k$, $\mathbf{b} = \bar{\mathbf{X}}_k$ and $\mathbf{l}' = (\underbrace{0, 0, \ldots, 0, 1}_{k}, 0, ..., 0)$. Then, the BLUP for this combination using available sample data is: [Henderson, 1975; Ghosh and Rao, 1994]

$$\widetilde{\mu}_{\bar{Y}_k} = \bar{\mathbf{X}}_k'\tilde{\beta} + \tilde{u} = \gamma_k\Big[\bar{y}_k + (\bar{\mathbf{X}}_k' - \bar{\mathbf{x}}_k')\tilde{\beta}\Big] + (1-\gamma_k)\bar{\mathbf{X}}_k'\tilde{\beta} \ . \tag{29}$$

To calculate the BLUP in equation (29), variance components are assumed to be known. Replacing the estimated values for the variance components in equation (29), 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 estimated for either individual or aggregated level analysis by the Fisher scoring algorithm, as presented in section 2.3.

An approximation for the Mean Square Error (MSE) of EPLUPs under a general LMM is: [Saei and Chambers, 2003b]

$$\mathcal{G}_1(\sigma) + \mathcal{G}_2(\sigma) + \mathcal{G}_3(\sigma) \ , \tag{30}$$

where:

$$\begin{aligned} \mathcal{G}_1(\sigma) &= (1-\gamma_k)\sigma_u^2 \\ \mathcal{G}_2(\sigma) &= (\bar{\mathbf{X}}_k - \gamma_k\bar{\mathbf{x}}_k)'\big[MSE(\tilde{\beta})\big](\bar{\mathbf{X}}_k - \gamma_k\bar{\mathbf{x}}_k) \\ \mathcal{G}_3(\sigma) &= \left(\frac{\sigma_e^2}{n_k}\right)^2\left(\sigma_u^2 + \frac{\sigma_e^2}{n_k}\right)^{-3} + \left[Var_\xi(\hat{\sigma}_u^2) + \frac{\sigma_u^4}{\sigma_e^4}Var(\hat{\sigma}_e^2) - 2\frac{\sigma_u^2}{\sigma_e^2}Cov(\hat{\sigma}_u^2, \hat{\sigma}_e^2)\right] \ , \end{aligned} \tag{31}$$

in which: $\sigma = (\sigma_u, \sigma_e)$. Detailed discussion of the MSE of EBLUPs is presented by Prasad & Rao (1990) and Saei & Chambers (2003a).

4 Contextual model

It is well known that estimation of 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. In SAE, it is common to use 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 values of the auxiliary variables can be included in the statistical population model as the contextual effect as in:

$$Y_{ik} = \mathbf{X}_{ik}^{*'}\beta^* + u_k^* + e_{ik}^* \quad ; \quad u_k^* \sim N(0, \sigma_{u^*}^2) \;\; \& \;\; e_{ik}^* \sim N(0, \sigma_{e^*}^2) \; . \tag{32}$$

Here, $\mathbf{X}_{ik}^{*'}$ involves both individual-level and aggregated-level covariates for ith unit within the kth area as below:

$$\mathbf{X}_{ik}^{*'} = [\mathbf{X}_{ik}' \mid \breve{\bar{\mathbf{X}}}_k'] \; , \tag{33}$$

where:

$$\breve{\bar{\mathbf{X}}}_k' = [\bar{X}_{k1} \quad \bar{X}_{k2} \quad \ldots \quad \bar{X}_{kP}] \; . \tag{34}$$

Note that, $\mathbf{X}_{ik}$ includes the intercept term, whereas $\breve{\bar{\mathbf{X}}}_k$ does not. The aggregated form of this population model is given as:

$$\bar{Y}_k = \bar{\mathbf{X}}_k'\beta^{**} + u_k^* + \bar{e}_k^* \quad ; \quad \bar{e}_k^* = \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} e_{ik}^* \sim N(0, \tfrac{\sigma_{e^*}^2}{N_k}). \tag{35}$$

Here,

$$\begin{aligned} &(\beta^{*I})' = [\beta_1^{*I} \quad \beta_2^{*I} \quad \ldots \quad \beta_P^{*I}] \; , \quad (\beta^{*c})' = [\beta_1^{*C} \quad \beta_2^{*C} \quad \ldots \quad \beta_P^{*C}] \; , \\ &\beta^{*'} = \Big[\beta_0^* \mid (\beta^{*I})' \mid (\beta^{*C})'\Big] \quad \& \quad \beta^{**'} = \Big[\beta_0^* \;\; (\beta_1^{*I} + \beta_1^{*C}) \;\; \cdots \;\; (\beta_P^{*I} + \beta_P^{*C})\Big]. \end{aligned} \tag{36}$$

Contextual models help researchers understand 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:

$$y_{ik} = \mathbf{X}^{*'}_{(s)ik}\beta^{*} + u^{*}_{k} + e^{*}_{ik} \tag{37}$$

and the true model for aggregate sample data is:

$$\bar{y}_{k} = \bar{\mathbf{X}}^{*'}_{(s)k}\beta^{**} + u^{*}_{k} + \bar{e}^{*}_{k} \tag{38}$$

where:

$$\mathbf{X}^{*'}_{(s)ik} = [\mathbf{x}'_{ik} \mid \breve{\bar{\mathbf{X}}}'_{k}] \quad \& \quad \bar{\mathbf{X}}^{*'}_{(s)k} = [\bar{\mathbf{x}}'_{k} \mid \breve{\bar{\mathbf{X}}}'_{k}]. \tag{39}$$

Note that, $\mathbf{X}^{*'}_{(s)ik}$ is the same as $\mathbf{X}^{*'}_{ik}$ when $i \in s$. The components of $\mathbf{X}^{*'}_{(s)ik}$ are the sample and population area level means.

If for some reasons the population data about the auxiliary variables are not available, we might replace the area population means by the corresponding sample means in the contextual model. Then an alternative working model would be:

$$y_{ik} = \mathbf{x}^{*'}_{ik}\beta^{*} + u^{*}_{k} + e^{*}_{ik} \tag{40}$$

Here, $\mathbf{x}^{*'}_{ik}$ included auxiliary information about the ith sample individual within the kth area as well as the kth area sample means, so:

$$\mathbf{x}^{*'}_{ik} = [\mathbf{x}'_{ik} \mid \breve{\bar{\mathbf{x}}}'_{k}] \quad \& \quad \breve{\bar{\mathbf{x}}}'_{k} = [\bar{x}_{k1} \quad \bar{x}_{k2} \quad \ldots \quad \bar{x}_{kP}] \; . \tag{41}$$

The aggregated form of this model presented in (40) is given as:

$$\bar{y}_k = \bar{\mathbf{x}}_k' \beta^{**} + u_k^* + \bar{e}_k^* \tag{42}$$

In aggregated-level analysis, the models presented in (18) and (42) are actually the same. This shows that the area-level models can involve existing contextual effects within the model, automatically using the sample instead of population.

5 Working Models

There are two population models considered in this paper as displayed in Table 1.

Table 1: Population Models

Population Model 1 (P_1):	$Y_{ik}^{(P_1)} = \mathbf{X}_{ik}' \beta + u_k + e_{ik}$
Population Model 2 (P_2):	$Y_{ik}^{(P_2)} = \mathbf{X}_{ik}^{*\prime} \beta^* + u_k^* + e_{ik}^*$

Population model *P1* is the standard unit-level model with random effects but not contextual effects. This model leads to model (5) for unit-level sample data and model (20) for aggregate sample data. In the current study we call these models, working model 1 (*W1*) and working model 2 (*W2*), respectively. One of the advantages of estimating the regression parameters using aggregate data is that area-level information can be used for covariates that were not included in the sample data but are available in the form of area population means. This leads to working model 3, (*W3*) as follows:

$$\bar{\mathbf{y}} = \bar{\mathbf{X}}' \beta + \epsilon \tag{43}$$

Population Model 2 (*P2*) incorporates contextual effects and leads to mode (37) for unit-level sample data and model (38) for aggregate sample data. We call these models working model 5 (*W5*) and working model 6 (*W6*), respectively, which both correctly use the population area level mean for the contextual part of the model. In practice, obtaining the population means of the covariates may be time consuming and in some situations it may be much easier to use the sample area level means in a unit-level contextual model, leading to working model 4 (W4), presented in (40). The working models discussed in this paper are presented in Table 2.

Table 2: Summary of Possible Working Models

	Working Models
W_1	$y_{ik}^{(W_1)} = \mathbf{x}_{ik}'\beta + u_k + e_{ik}$
W_2	$\bar{y}_k^{(W_2)} = \bar{\mathbf{x}}_k'\beta + u_k + \bar{e}_k$
W_3	$\bar{y}_k^{(W_3)} = \bar{\mathbf{X}}_k'\beta + u_k + \bar{e}_k$
W_4	$y_{ik}^{(W_4)} = \mathbf{x}_{ik}^{*\prime}\beta^* + u_k^* + e_{ik}^*$
W_5	$y_{ik}^{(W_5)} = \mathbf{X}_{(s)ik}^{*\prime}\beta^* + u_k^* + e_{ik}^*$
W_6	$\bar{y}_k^{(W_6)} = \bar{\mathbf{X}}_{(s)ik}^{*\prime}\beta^* + u_k^* + e_{ik}^*$

The six working models can be characterised as follows:

- W_1: A unit-level model without considering any contextual effects.
- W_2: An area-level model which involves the sample area means in the model as the auxiliary information.
- W_3: An area-level model which involves the population area means in the model as the auxiliary information.
- W_4: A unit-level model which involves the sample area means in the model as possible contextual effects.
- W_5: A unit-level model which involves the population area means in the model as possible contextual effects.
- W_6: A area-level model which involves both sample and population area means.

The expectation of the regression parameters estimations associated with each working model can be obtained under both population model.

When P_1 is the true population model:

- W_1 is the correct unit-level model under P_1 leading to unbiased estimates.
- W_2 is the correct area-level model under P_1 leading to unbiased estimates, but with larger variances than those estimated using W_1, because of the use of aggregate data.

- Estimates based on W_3 are biased under P_1 and the bias term is due to the difference between the area population means and area sample means.
- For W_4, W_5 and W_6, the regression parameter estimates are unbiased but these contextual models are inefficient due to over-fitting of model parameters.

When P_2 is the true population model:

- For W_1, model parameter estimates are biased due to omission of the existing contextual effects in P_2.
- For W_2, the resulting estimates are slightly biased as W_2 does not include area population means, but implicitly includes sample area means.
- For W_3, the resulting estimates are slightly biased as W_3 does not include area sample means.
- For W_4, the parameter estimates are slightly biased and the bias term is due to the difference between area sample and population means.
- W_5 is the correct unit-level model under P_2 leading to unbiased estimates.
- W_6 is the correct area-level model under P_2 leading to unbiased estimates, but the co-linearity between sample and population area means is an issue to be considered in this case.

For each working model we can consider the associated synthetic estimation and EBULP given by (27) and (29).

6 An Empirical Study

This section presents the results of a model-assisted design-based simulation study to empirically assess the bias and Mean Square Error (MSE) of synthetic estimators and EBLUPs based on the unit-level and area-level working models discussed in section 5. As an example, we suppose that there is an 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 the simulation presented here, population data is generated based on two different population models, separately as presented in Table 1.

Parameter values for the population models of the relation between weekly gross salary and hours worked for individuals over 15 were obtained from the Australian Australian 2006 Census. Table (4) presents the model parameter values used in generating the populations of individuals. Sample units are then selected based on a stratified random sampling design in which the sample sizes in the 57 areas are allocated proportionally to their population sizes. The six working models presented in Table 2 are then fitted on the sample data in order to compare the resulting estimates based on these models. A total sample

Table 3: Parameter Values Considered in the Population Models

Population Model 1			
$\beta' = [\beta_0 \ \ \beta_1]$	σ_u	σ_e	λ
[322.45 14.93]	114.3530	384.6394	0.0884
Population Model 2			
$\beta' = [\beta_0^* \ \ \beta_1^{*I} \ \ \beta_1^{*C}]$	σ_u^*	σ_e^*	λ^*
[-123.6008 14.93 3.7724]	114.3530	384.6394	0.0884

of 2133 was used and the resulting sample sizes varied from 1 to 398 with an average of 37. The details of the sample allocation are given in Appendix 1 (Table 9) .

The estimation techniques in this study were evaluated by calculating the relative Root Mean Squared Error (rRMSE) for each area using the different working models as follows:

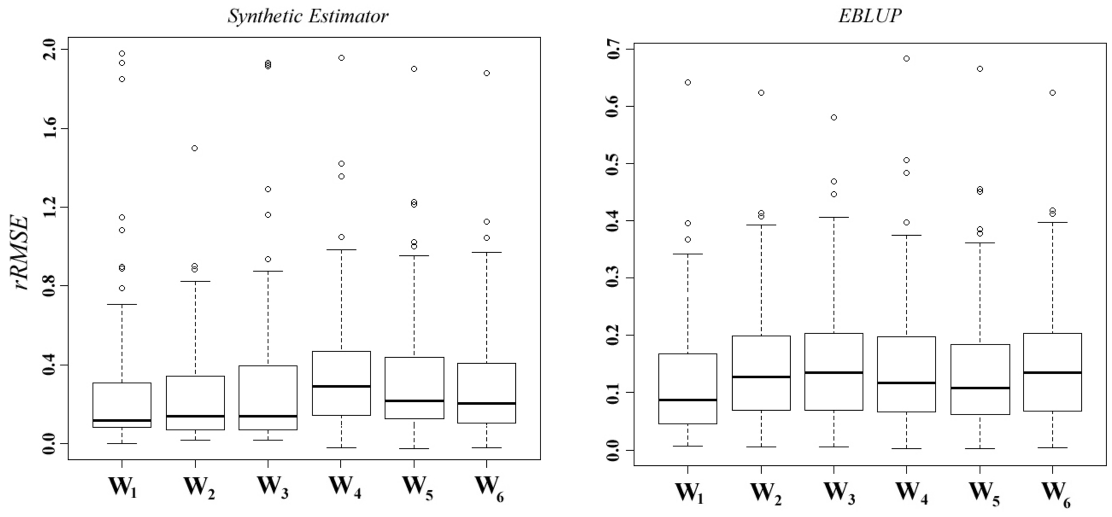
$$rRMSE_k = \frac{\sqrt{MSE(\hat{\hat{\bar{Y}}}_k)}}{\bar{Y}_k} \quad ; \quad k = 1, \ldots, 57 \tag{44}$$

where,

$$MSE(\hat{\hat{\bar{Y}}}_k) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{\hat{\bar{Y}}}_{k(m)} - \bar{Y}_k\right] \tag{45}$$

Note that a list of $M = 1000$ samples were selected in this study. Here, $\hat{\hat{\bar{Y}}}_{k(m)}$ is the estimate of the kth area mean based on mth sample. Using side by side box plots, Figure 1 and 2 show the resulting rRMSEs for the synthetic estimates and EBLUPs obtained based on six working models (presented in Table 2), considering two working models (presented in Table 1) using the 1000 samples selected.

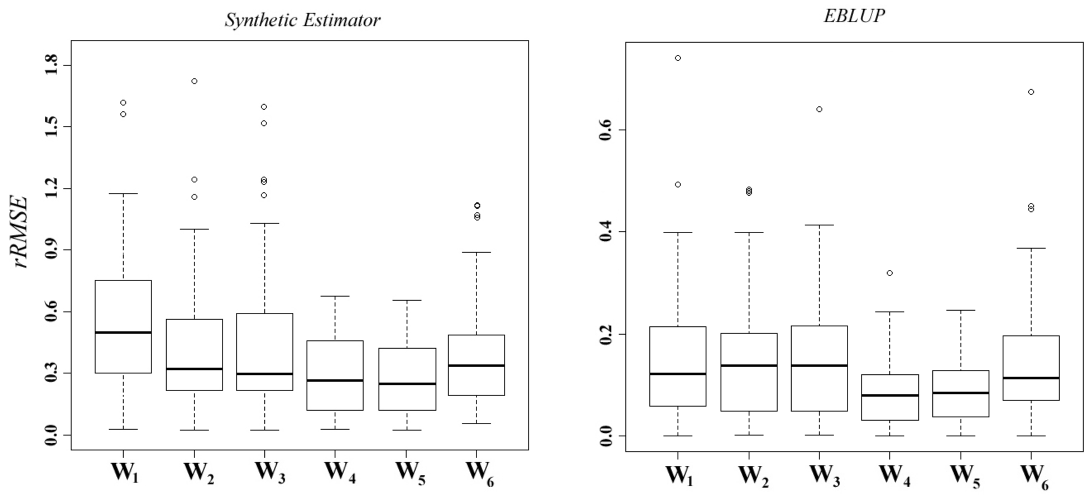
Figure 1: rRMSE under *Population 1*



As can be seen in Figure 1, for synthetic estimation, using W_1 leads to the smallest mean (and median) rRMSE for 57 areas and smallest deviation in the rRMSE. However, the performance of W_2 and W_3 in terms of rRMSE is not greatly worse. Using W_4, W_5 or W_6 which allow for contextual effects have noticeable larger mean and median rRMSE. In particular W_4 has the worse performance. Looking at the results for the EBLUPs in Figure 1, the resulting estimates using W_1 performs much better than any other estimates. Use of W_5 produced EBLUPs with similar average and median rRMSE. Use of the EBLUP technique leads to considerable gains related to the synthetic estimates in terms of rRMSE for W_1, W_4, and W_5. On the other hand, the average rRMSE increases for W_2 and W_3 compared with corresponding synthetic estimates.

In figure 2, P_2 is considered as the true population model. For the synthetic method, resulting estimators using the area-level working model W_2 are better in terms of rRMSE than those calculated based on unit-level model W_1. The best approach is to fit unit-level contextual working models using either the sample or population area means as the area-level or contextual effects, as in W_4 and W_5. However, use of the EBLUP technique seems to correct much of the problem with W_1. As can be seen, EBLUP estimates based on W_1 and W_2 have similar properties under P_2 in terms of rRMSE. However, using W_4 or W_5 leads to the best estimators in such a case, while W_6 performs better than W_1, W_2 and W_3.

Figure 2: rRMSE under *Population 2*



Assuming P_2 applies for the population, fitting working model W_1 leads to biased parameter estimates. For the aggregate data the true sampling model is the one presented in (38). Therefore, 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 (37) and (38). Therefore, the dimension reduction in calculating model parameter estimates is an advantage of applying W_2.

The relative performance of the different working models can be examined by looking at the mean of the Root MSE, as summarised in Table 4.

Table 4: Mean of Empirical Root MSE over areas and 1000 simulations averaged over 57 areas

			Syn. Est.		**EBLUP**	
Working Model	Level	Contextual Effect	P_1	P_2	P_1	P_2
W_1	Unit	None	76.1	111.4	61.3	91.3
W_2	Agg	Sample	81.9	79.9	92.3	90.1
W_3	Agg	Pop	84.1	80.3	91.1	91.1
W_4	Unit	Sample	93.8	54.0	80.9	82.3
W_5	Unit	Pop	92.4	53.9	71.2	82.8
W_6	Agg	Pop	93.3	78.6	92.9	89.3

As can be seen in Table 4, for $P1$, *i*) W_1 seems to be the best choice for both synthetic estimation technique and EBLUP. *ii*) $W2$ is not a lot worse that $W1$ for synthetic estimation but it is for EBLUP. *iii*) EBLUPs are better than synthetic estimators for the unit-level models but not for the aggregated-level models. *iv*) allowing for contextual effects makes things worse for synthetic estimators and EBLUPs in terms of root MSE. For P2, *i*) $W1$ is the worst choice considering the synthetic estimation method but the estimation results are improved by using EBLUP. *ii*) Unit-level models with the contextual effects perform best for synthetic estimations and EBLUPs, while EBLUPs have larger root MSEs. Something is going on with the EBLUPs through estimation of variance components when adding the contextual effects in the working models. *iii*) Using sample means as contextual effects is as good as using population means.

If we are restricted to using regression synthetic estimates, then perhaps $W2$ is a reasonable compromise choice. If EBLUP approach is to be used, then $W1$ or $W5$ is a reasonable choice. I would be noted that, estimation results depend on the strength of contextual effects. The difference between parameter estimates using $W1$ and $W2$ may be due to other omitted variables and the effect of aggregation on the regression parameters relating these omitted variables and the included covariates.

Here, the properties of the resulting estimates using W_1 and W_2 are examined when P_2 is the true population model. These two models are those that are most commonly considered and will examine the properties of the resulting estimation using these models in more details. Considering the area means as the main targets of inference, the bias of the unit- and area-level synthetic estimate under P_2 are:

$$\begin{aligned} Bias_{\xi_{(P_2)}}\left(\widehat{\bar{Y}}_k^{SU}\right) &= \bar{\mathbf{X}}_k' E_{\xi_{(P_2)}}[\hat{\beta}^{(W_1)} - \beta^{**}] \ , \\ Bias_{\xi_{(P_2)}}\left(\widehat{\bar{Y}}_k^{SA}\right) &= \bar{\mathbf{X}}_k' E_{\xi_{(P_2)}}[\hat{\beta}^{(W_2)} - \beta^{**}] \ . \end{aligned} \tag{46}$$

The subscript ξ denotes the expectation, bias, MSE and variance under the assumed population model. It can be shown that $E_{\xi_{(P_2)}}\left[\hat{\beta}^{(W1)}\right] \approx \beta^{*I}$ and $E_{\xi_{(P_2)}}(\hat{\beta}^{(W1)} - \beta^{**}) \approx [0 \quad \beta^{*C}]'$. Therefore, the bias of the unit-level synthetic estimator for kth area mean is $\bar{\mathbf{X}}_k \beta^{*}$. For $\hat{\beta}^{(W2)}$, 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 $\hat{\beta}^{(W2)}$ to be a less biased estimate of β^{**} than $\hat{\beta}^{(W1)}$.

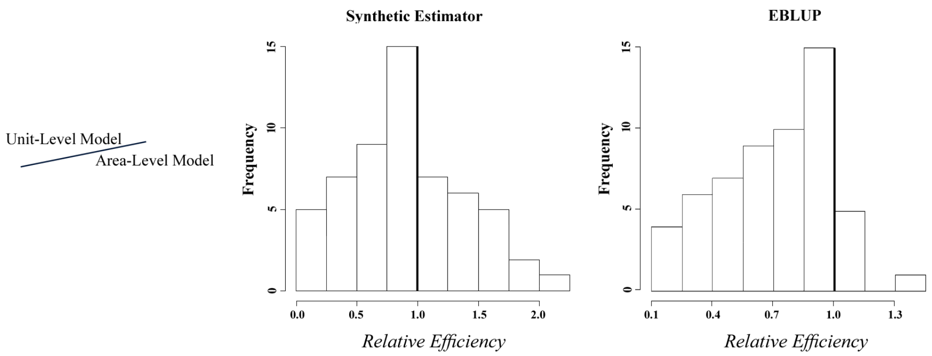
The bias of the unit-level EBLUP for kth area mean is calculated as follows:

$$Bias_{\xi}\Big(\widetilde{\widetilde{Y}}_k^{(W_1)}\Big) = \Big[\bar{\mathbf{X}}'_k - E_{\xi}(\widehat{\gamma}_k)\bar{\mathbf{x}}'_k\Big] E_{\xi}\Big(\tilde{\beta}^{(W_1)} - \beta^{**}\Big) + Cov_{\xi}\Big[\widehat{\gamma}_k, \big(\bar{y}_k - \bar{\mathbf{x}}'_k\tilde{\beta}^{(W_1)}\big)\Big]. \quad (47)$$

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 3 summarizes the empirical results by giving the ratio of MSEs for the SAEs based on unit-level model ($W1$) and area-level model ($W2$) 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 3: The Relative Efficiency of Unit-level Model to Area-level Model

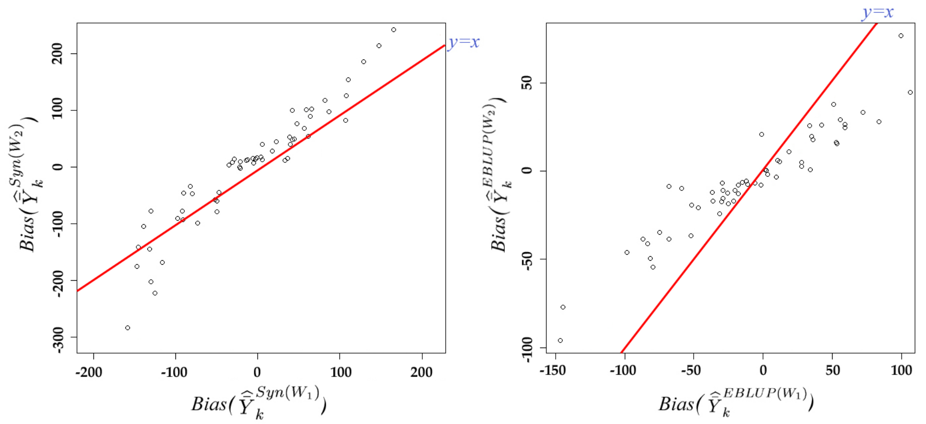


A comparison between the resulting bias based on the synthetic estimation approach and EBLUP technique is presented in Figure 3 for the target areas. For positive biases of the synthetic estimates, unit-level and area-level results look similar in terms of bias values. However, when the resulting biases for unit-level synthetic estimates are negative, less biased synthetic estimates can be calculated based on area-level model ($W2$). 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 the unit-level synthetic estimator and EBLUP for kth area mean are:

$$\begin{aligned} Bias_{\xi}\left(\tilde{\tilde{Y}}_k^{(W_1)}\right) &\approx (1-\gamma_k)\bar{\mathbf{X}}_k' Bias_{\xi}\left(\tilde{\beta}^{(W_1)}\right) = \left(\frac{\frac{\sigma_e^2}{n_k}}{\sigma_u^2+\frac{\sigma_e^2}{n_k}}\right)\bar{\mathbf{X}}_k' Bias_{\xi}\left(\tilde{\beta}^{(W_1)}\right) \\ Bias_{\xi}\left(\tilde{\tilde{Y}}_k^{(SU)}\right) &\approx \bar{\mathbf{X}}_k' Bias_{\xi}\left(\tilde{\beta}^{(W_1)}\right). \end{aligned} \tag{48}$$

As shown in (48), 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 4: Resulting Bias for Synthetic Estimators and EBLUPs



Means and variances of the parameter estimates (including the variance components estimated for calculating the EBLUPs) using working models used in this numerical study are presented in Table 5. 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_1.

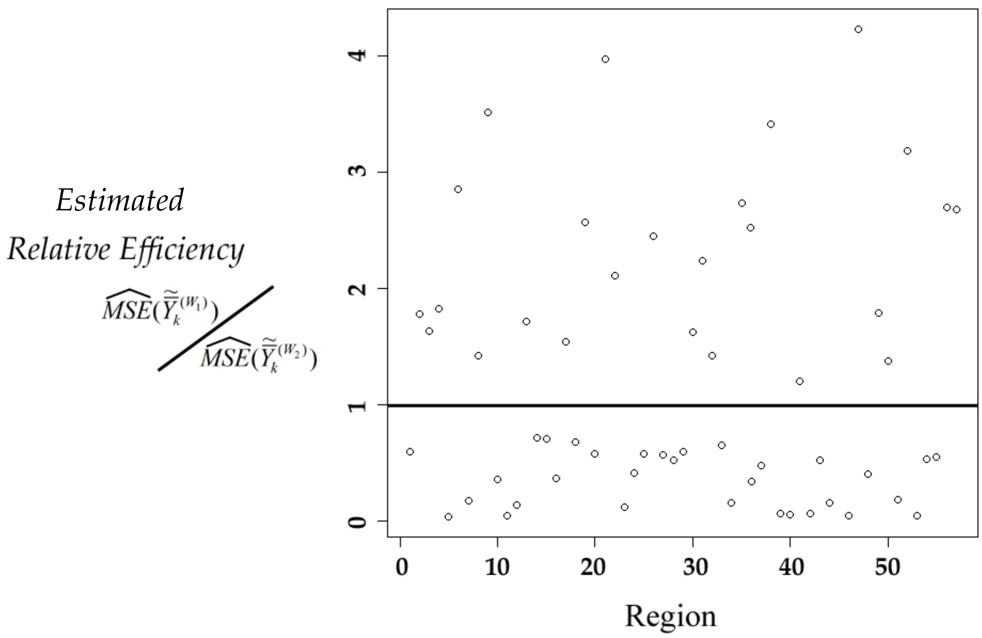
Table 5: Parameter Estimates under Population 2

	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

In the simulation presented in this section, MSEs have been calculated by the simulation. In real situations the data would come from surveying the target population and the required MSEs will be estimated. Then, the equation presented in (31) can be used in order to estimate the MSE of resulted predictions. Figure 5 shows the estimated relative efficiency for 57 area EBLUPs based on W_1 over W_2 under P_2. As can be seen in Figure 5, the resulting area-level EBLUPs calculated based on W_2 have smaller estimated MSEs in many areas.

As can be seen in Figure 5 the estimated EBLUPs calculated based on W_1 comparing with those calculated based on W_2 have smaller estimated MSEs for some areas and have larger estimated MSEs for some others. If similar results are obtained in practice, this can be a sign of possible area-level or contextual effects to be present in the actual population

Figure 5: Estimated Relative Efficiency for EBLUPs based on W_1 over W_2 under P_2



model. Based on previous discussions, W_2 can be fitted on the sample data leading to reasonably precise estimates in terms of estimated MSE, when area means are the main targets of inferences while the matrix dimensions in W_1 calculating required estimates are much less than those in W_2. This may make W_2 to be preferred in practice.

7 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 aggregated-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.

We need to be careful about area effects related to contextual variables, as random effects do not account for these. Based on the discussions presented in this paper, the presence of contextual effects can be assessed by i) comparing parameter estimates arising from $W1$ and $W2$, ii) fitting $W4$, which uses sample area means iii) fitting $W5$, which uses population area level means. If $P1$ seems to apply, then use $W1$, preferably using EBLUP. If $P2$ seems to apply, then use regression synthetic technique based on $W5$ or $W4$. 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. (2009). Model-based Approach to Small Area Estimation. Chapter 32 in

C.R. Rao and D. Pfeffermann (eds.) *Handbook of Statistics.* **29(B)**, 251-288. *Sample Surveys: Theory, Methods and Inference.* Elsevier; Amsterdam, North Holland.

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.

Lehtonen, R., and Veijanen, A. (2009). Design-based Methods of Estimation for Domains and Small Areas. Chapter 31 in C.R. Rao and D. Pfeffermann (eds.) *Handbook of Statistics.* **29(B)**, 219-249. *Sample Surveys: Theory, Methods and Inference.* Elsevier; Amsterdam, North Holland.

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.

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

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.

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.

Appendix

Table 6: 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 7: Weekly Gross Salary

STATE	No.	Statistical Subdivisions	Income (Mean)	Std. Dev.	N	Std. Err. Mean
ACT	1	Canberra	963.72045	836.09364	229557	1.7450571
NSW	2	Murray	566.09301	24932.468	20105	3.720431
	3	Northern	573.95819	549.74003	115688	1.6162674
	4	Murrumbidgee	606.5969	552.63812	97902	1.766221
	5	Sydney	835.26184	831.12165	2699536	0.505848
	6	Richmond-Tweed	545.82145	521.12234	152499	1.3344619
	7	South Eastern	653.21868	633.20965	135506	1.7201577
	8	Central West	610.23421	591.73768	114364	1.7497845
	9	Mid-North Coast	511.63105	489.75066	198991	1.0978887
	10	Illawarra	644.48308	645.54542	268424	1.2459945
	11	Hunter	624.19457	642.47515	408379	1.005367
	12	Far West	546.21759	552.16024	14964	4.5137893
	13	North Western	592.19592	572.21542	72193	2.1296685
NT	14	Northern Territory - Bal	636.95568	675.38653	49167	3.0458993
	15	Darwin	855.21733	688.49003	66787	2.6641071
QLD	16	Brisbane	746.18657	705.6472	1193749	0.6458492
	17	Central West	653.14114	595.454117	7163	7.0355956
	18	Far North	643.69176	587.31543	148088	1.5262079
	19	South West	655.83141	598.83141	16021	4.7281915
	20	Fitzroy	749.75586	740.05698	121241	2.1253987
	21	Moreton	540.54207	483.38593	32745	2.6712929
	22	North West	852.6017	761.7846	18142	5.6557422
	23	Mackay	818.56413	812.15257	95597	2.6267304
	24	Wide Bay-Burnett	516.11945	494.0813	173635	1.1857135
	25	Northern	697.82914	632.53831	130340	1.752056
	26	Darling Downs	601.26056	562.16546	143547	1.4837718
SA	27	Adelaide	659.51368	629.83834	786097	0.7103805
	28	Outer Adelaide	614.42725	568.4696	85614	1.9428302
	29	Northern	600.54169	587.54169	50536	2.613131
	30	Murray Lands	524.94057	475.45577	46271	2.2103227
	31	Eyre	587.70572	547.67934	22360	3.6626082
	32	Yorke and Lower North	515.84562	484.75968	31261	2.7417324
	33	South East	612.26209	556.75698	29581	3.2371233
TAS	34	Northern	565.56349	525.92225	93276	1.7220136
	35	Greater Hobart	643.08777	598.88874	140360	1.5985435
	36	Mersey-Lyell	546.35121	504.07257	74278	1.8495368
	37	Southern	512.5	447.13588	24060	3.0760562
VIC	38	Melbourne	750.5854	748.62431	2416087	0.4816235
	39	Central Highlands	589.57634	554.42658	97166	1.7786352
	40	Ovens-Murray	599.27068	534.71235	64341	2.1080277
	41	Gippsland	596.30416	585.77684	107966	1.7827428
	42	Goulburn	582.17638	530.10244	130811	1.465675
	43	Mallee	544.57239	491.10582	59294	2.0168319
	44	Loddon	597.91793	578.27649	115318	1.7028915
	45	Barwon	633.6784	615.37539	177890	1.4590305
	46	Wimmera	555.73123	511.61537	33806	2.782538
	47	Western District	611.63216	588.14523	67567	2.2626494
	48	East Gippsland	567.13903	571.55861	54990	2.4373557
WA	49	Lower Great Southern	605.56038	585.84961	35110	3.019516
	50	Perth	785.10057	770.44457	975121	0.7802111
	51	Pilbara	1297.373	1102.2988	22259	7.388337
	52	South West	660.28024	672.58787	138739	1.8057137
	53	South Eastern	896.80946	837.11511	31468	4.7190069
	54	Upper Great Southern	637.20282	586.34493	11651	5.4321478
	55	Central	680.06813	640.86909	36182	3.3691709
	56	Kimberley	694.42033	666.30042	16820	5.1375622
	57	Midlands	641.67349	612.86733	32967	3.3754117

Table 8: Hours Worked

STATE	No.	Statistical Subdivisions	Hours Worked (Mean)	Std. Dev.	N	Std. Err. Mean
ACT	1	Canberra	38.535109	19.013724	164616	0.046831
NSW	2	Murray	43.046029	22.485444	12351	0.2023254
	3	Northern	41.348493	21.544124	67769	0.0827586
	4	Murrumbidgee	41.176176	20.973207	62097	0.0841646
	5	Sydney	40.357063	19.686667	1784299	0.014738
	6	Richmond-Tweed	37.085496	20.199353	80688	0.0711104
	7	South Eastern	39.892999	20.630821	81728	0.0721657
	8	Central West	40.926679	21.169946	68384	0.0809548
	9	Mid-North Coast	36.814685	20.119273	95702	0.0650357
	10	Illawarra	37.762631	19.707078	149514	0.0509661
	11	Hunter	38.433551	19.976324	230982	0.0415649
	12	Far West	41.230153	21.998809	7432	0.2551798
	13	North Western	41.996767	21.555434	43617	0.1032117
NT	14	Northern Territory - Bal	42.058692	21.149081	29595	0.1229369
	15	Darwin	43.780662	19.654855	500078	0.0878307
QLD	16	Brisbane	40.133727	19.846652	810831	0.0220406
	17	Central West	47.562737	22.054888	5268	0.3038659
	18	Far North	41.655183	20.601954	100001	0.0651488
	19	South West	47.034528	22.646135	11831	0.2082013
	20	Fitzroy	43.382941	21.363771	81809	0.0746826
	21	Moreton	41.056225	21.196183	19822	0.1505511
	22	North West	47.656069	21.294848	13494	0.1833176
	23	Mackay	44.777687	21.519457	67956	0.0825501
	24	Wide Bay-Burnett	39.970465	20.872827	88862	0.0700202
	25	Northern	42.418276	20.825458	87826	0.0702721
	26	Darling Downs	41.702297	21.335674	91064	0.0707022
SA	27	Adelaide	38.097493	19.203949	477231	0.0277988
	28	Outer Adelaide	39.303404	20.698605	52824	0.0900586
	29	Northern	41.083465	21.483465	28449	0.1273959
	30	Murray Lands	40.628643	21.005359	27417	0.1268587
	31	Eyre	41.524671	22.051422	14288	0.1844807
	32	Yorke and Lower North	40.984263	22.236633	16363	0.1738351
	33	South East	39.90694	20.484917	18762	0.1495528
TAS	34	Northern	38.136817	19.935228	53050	0.0865523
	35	Greater Hobart	37.042095	18.737503	83633	0.0647922
	36	Mersey-Lyell	39.074302	20.224338	40369	0.1006585
	37	Southern	37.797148	20.333716	12832	0.1795021
VIC	38	Melbourne	39.408675	19.757399	1580782	0.0157866
	39	Central Highlands	38.504711	20.303516	58162	0.0841883
	40	Ovens-Murray	39.762735	20.649661	40775	0.1022624
	41	Gippsland	39.116529	21.027399	62092	0.0843855
	42	Goulburn	40.592672	21.107216	80213	0.0745261
	43	Mallee	41.084178	20.919517	35793	0.1105739
	44	Loddon	38.625031	20.739254	68439	0.0792759
	45	Barwon	37.945018	20.035356	106835	0.0612972
	46	Wimmera	40.984469	21.557419	20218	0.1516099
	47	Western District	40.634364	21.828231	42158	0.103111
	48	East Gippsland	39.545577	21.725615	30355	0.1246973
WA	49	Lower Great Southern	41.378171	22.002367	21682	0.1494238
	50	Perth	39.746568	20.361196	656483	0.0251299
	51	Pilbara	49.725775	21.904649	17905	0.1637002
	52	South West	40.374651	21.257887	83523	0.073558
	53	South Eastern	47.308024	22.966973	23292	0.1504875
	54	Upper Great Southern	47.166524	23.523255	8164	0.260343
	55	Central	43.531958	22.095074	23343	0.1446163
	56	Kimberley	42.819141	21.710507	11755	0.2002436
	57	Midlands	45.17157	22.916712	21210	0.1573555

Table 9: 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	