SMALL AREA ESTIMATION UNDER INFORMATIVE SAMPLING

F. Verret[1], M.A. Hidiroglou[2] and J.N.K Rao[3]

ABSTRACT

Population unit level models are often used in model-based small area estimation of totals and means but the models may not hold for the sample if the sampling design is informative. As a result, standard methods, assuming the model holds for the sample, can lead to biased estimators. We propose to study alternative methods that use the survey weight as an additional auxiliary variable in the sample model and/or in the estimation of means and MSEs using the pseudo-EBLUP approach proposed by You and Rao (2002). We report the results of a simulation study on the bias and the MSE of the proposed point estimators and on the relative bias of the MSE estimators, using informative sampling schemes to generate the samples.

KEY WORDS: Informative Sampling, Small Area Estimation, Unit-Level Model.

RÉSUMÉ

Les modèles au niveau des unités de population sont souvent utilisés en estimation pour petits domaines reposant sur des modèles pour des totaux et des moyennes. Ces modèles peuvent ne pas être applicables à l'échantillon si le plan d'échantillonnage est informatif. Les méthodes habituelles, supposant que le modèle est approprié pour l'échantillon, peuvent donc mener à des estimateurs biaisés. Nous proposons d'étudier des méthodes alternatives utilisant les poids de sondage comme variable auxiliaire supplémentaire dans le modèle ajusté à l'échantillon et/ou dans l'estimation des moyennes et EQMs en utilisant l'approche pseudo-EBLUP proposée par You et Rao (2002). Nous présentons les résultats d'une étude de simulation sur le biais et l'EQM des estimateurs ponctuels proposés et sur le biais relatif des estimateurs d'EQM, lorsque les échantillons sont obtenus à partir de plans informatifs.

MOTS CLÉS : Échantillonnage informatif; estimation pour les petites régions; modèle au niveau des unités.

1. INTRODUCTION

Estimates of population parameters, such as means and totals, are often required for small areas (domains). Those parameters can be estimated using direct estimation which involves survey weights that incorporate adjustments due to non-response and / or auxiliary data. However, if the target domain realized sample size is too small or even zero, then the area-specific direct estimators are not reliable or not feasible and it is necessary to use indirect estimators based on models that provide a link to related areas and thus borrow strength across areas (Rao, 2003). Linking models are either defined at the unit level (when data are available at the unit level) or at the small area level (when data are available at the small area level or data are available at the unit level aggregated to the small area level). The small areas can be identified before or after sampling takes place.

In this paper, we consider small areas that are defined prior to the sampling process as strata or clusters, and data that are available at the unit level. Unit level models are used to borrow strength across related small areas to build efficient small area estimators. A potential problem with this set-up is that the sample design could be informative. That is, the unequal probabilities associated with the selection of the small areas may possibly be related to the true area means, and, the unequal probabilities associated with the selection of units within the sampled small areas may also be related to their outcome values. This will imply that a model holding for the population values no longer holds for the sample data. This in turn implies that the effects of the sampling process on the distribution of the observed outcomes may bias the inference very severely. Pfeffermann and Sverchkov (2007) noted that this effect could be controlled by either including all of the design variables used for sample selection among the model covariates, or using the survey weight as a surrogate. However, they viewed both solutions as not practical: the design variables might not be available at the inference stage, and the survey weights might not be available for all the population units. Pfeffermann and Sverchkov (2007) proposed a

[1] F. Verret, Statistics Canada, 16 J, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6, Francois.Verret@statcan.gc.ca
[2] M.A. Hidiroglou, Statistics Canada, 16 D, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6, Mike.Hidiroglou@statcan.gc.ca
[3] J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6, jrao@math.carleton.ca

method to reduce sample selection bias by taking into account the relationship between the sampling weight and the components of the unit level model.

In this paper we will compare, via simulations, the unweighted empirical best linear unbiased predictor, EBLUP, and the weighted pseudo-EBLUP to an estimator proposed by Pfeffermann and Sverchkov (2007). We consider a two-stage sampling design where all the first stage units (small areas) are selected with certainty and the second stage units are selected using the Rao-Sampford πPS sampling scheme (Rao 1965, Sampford 1967). We include or exclude the second stage weights as additional auxiliary variables for the models used to construct the EBLUP and the pseudo-EBLUP small area estimators. The unit level small area procedures are summarized in section 2. The simulation study and its results are described in section 3. Concluding remarks are provided in section 4.

2. SMALL AREA PROCEDURES CONSIDERED

We first describe the two-stage sampling design that we will use. A survey population *U* is first split into *M* distinct and non-overlapping small areas, with N_i elements in area i. A sample *s* is then selected from *U* in two stages as follows. First, a sample of *m* areas is selected from the *M* areas using a *πPS* scheme. The first stage probability associated with the *i*-th area is denoted as $\pi_i, i = 1,\text{K},M$. Secondly, a sample of n_i elements is selected from the N_i elements within the selected areas with probability $\pi_{j|i}$, $i = 1,\text{K},m$ and $j = 1,\text{K},N_i$. The first stage sampling weight is $w_i = 1/\pi_i$ and the second-stage weight is $w_{j|i} = 1/\pi_{j|i}$. The overall weight for a unit *j* within small area *i* is $w_{ij} = w_i w_{j|i}$. In this paper, we assume that all the areas are sampled, i.e., *m*=*M* and $\pi_i = 1$, and we assume N_i is large ($i = 1,...,M$).

We briefly describe how the unweighted EBLUP, the weighted pseudo-EBLUP, and the estimator proposed by Pfeffermann and Sverchkov (2007) are computed. Denote the variable of interest as *y*, and the auxiliary variables as $\boldsymbol{x}$. Consider the following hierarchical (or nested error) model given by

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}; \qquad j = 1,...,N_i; \qquad i = 1,...,M \tag{2.1}$$

where $v_i \overset{\text{iid}}{:} N\left(0,\sigma_v^2\right)$ and $e_{ij} \overset{\text{iid}}{:} N\left(0,\sigma_e^2\right)$. The variance components σ_v^2 and σ_e^2 are estimated from the sample using restricted maximum likelihood (REML) ignoring the design informativeness (i.e., assuming the model holds for the sample) resulting in the estimators $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ (see Rao 2003, pp. 100-102). Under model (2.1), the *i*-th area mean $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i$ may be approximated by $\theta_i = \bar{\mathbf{X}}_i' \beta + v_i$.

The unweighted EBLUP estimator (Battese, Harter and Fuller, 1988) of the *i*-th area mean θ_i is given by

$$\hat{\theta}_{iu} = \hat{\gamma}_i \bar{y}_i + \left(\bar{\boldsymbol{X}}_i - \hat{\gamma}_i \bar{\boldsymbol{x}}_i\right)^T \hat{\boldsymbol{\beta}}_u \tag{2.2}$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_i\right)$, $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} \Big/ n_i$ is the unweighted sample mean of the response variable *y*, $\bar{\boldsymbol{x}}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \Big/ n_i$, $\bar{\boldsymbol{X}}_i$ is the vector of known population means of the $\boldsymbol{x}_{ij}$'s for the *i*-th area and $\hat{\boldsymbol{\beta}}_u$ is the estimated unweighted regression vector given by

$$\hat{\boldsymbol{\beta}}_u = \left\{ \sum_{i=1}^{M} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \left(\boldsymbol{x}_{ij} - \hat{\gamma}_i \bar{\boldsymbol{x}}_i \right)^T \right\}^{-1} \left\{ \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left(\boldsymbol{x}_{ij} - \hat{\gamma}_i \bar{\boldsymbol{x}}_i \right) y_{ij} \right\}. \tag{2.3}$$

Letting $\tilde{w}_{ij} = w_{ij} \Big/ \sum_{k=1}^{n_i} w_{ik}$ denote the normalized weights for area *i*, the weighted pseudo-EBLUP estimator (You and Rao, 2002) is given by

$$\hat{\theta}_{iw} = \hat{\gamma}_{iw}\,\bar{y}_{iw} + \left(\bar{\boldsymbol{X}}_i - \hat{\gamma}_{iw}\bar{\boldsymbol{x}}_{iw}\right)^T \hat{\boldsymbol{\beta}}_w \tag{2.4}$$

where $\hat{\gamma}_{iw} = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + \delta_i^2 \hat{\sigma}_e^2\right)$ with $\delta_i^2 = \sum_{j=1}^{n_i} w_{ij}^2$, $\bar{y}_{iw} = \sum_{j=1}^{n_i} w_{ij} y_{ij}$ and $\bar{\boldsymbol{x}}_{iw} = \sum_{j=1}^{n_i} w_{ij}\boldsymbol{x}_{ij}$ are the i-th area weighted means of y and $\boldsymbol{x}$, and

$$\hat{\boldsymbol{\beta}}_w = \left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i} w_{ij}\boldsymbol{x}_{ij}\left(\boldsymbol{x}_{ij} - \hat{\gamma}_{iw}\bar{\boldsymbol{x}}_{iw}\right)^T\right\}^{-1} \left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i} w_{ij}\left(\boldsymbol{x}_{ij} - \hat{\gamma}_{iw}\bar{\boldsymbol{x}}_{iw}\right) y_{ij}\right\}. \tag{2.5}$$

Pfeffermann and Sverchkov (2007) assumed that the sampling weights $w_{j|i}$ satisfy

$$E_{si}\left(w_{j|i}\middle|\,\mathbf{x}_{ij}, y_{ij}, v_i, i \in s\right) = k_i \exp\left(\mathbf{x}_{ij}^T\mathbf{a} + b y_{ij}\right), \tag{2.6}$$

where $k_i = N_i n_i^{-1} \sum_{j=1}^{N_i} \exp\left(-\boldsymbol{x}_{ij}^T\boldsymbol{a} - b y_{ij}\right) \Big/ N_i$, and $\boldsymbol{a}$ and b are fixed unknown constants. Under this assumption, they obtained an estimator of $\bar{Y}_i$ that provides protection against informative sampling. It is given by

$$\hat{\bar{Y}}_{i,ps} = N_i^{-1}\left[\left(N_i - n_i\right)\hat{\theta}_{iu} + n_i\left\{\bar{y}_i + \left(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{x}}_i\right)^T \hat{\boldsymbol{\beta}}_u\right\} + \left(N_i - n_i\right)\hat{b}\,\hat{\sigma}_e^2\right]. \tag{2.7}$$

The last term of (2.7) corrects the EBLUP estimator for any bias due to informative sampling under (2.6). Pfeffermann and Sverchkov (2007) obtained an estimator of b by regressing the sampling weights $w_{j|i}$ on $k_i \exp\left(\boldsymbol{a}^T\boldsymbol{x}_{ij} + b\,y_{ij}\right)$. The coefficients k_i, $\boldsymbol{a}$ and b may be estimated by fitting the model (2.6) using procedures REG and NLIN in SAS. This involves iterative calculations and the initial values for $\boldsymbol{a}$ and b are obtained by regressing $\log(w_{j|i})$ on $\boldsymbol{x}_{ij}$ and y_{ij}, and the initial values for $\hat{k}_q, q = 1,...,m$ are taken as $k_q = N_q / n_q$. The estimator (2.7) may be approximated for large N_i by

$$\hat{\theta}_{i,ps} = \hat{\theta}_{iu} + \hat{b}\hat{\sigma}_e^2. \tag{2.8}$$

The mean squared error (MSE) of the EBLUP estimator $\hat{\theta}_{iu}$ is estimated by

$$mse\left(\hat{\theta}_{iu}\right) = g_{1i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + 2g_{3i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) \tag{2.9}$$

where

$$g_{1i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) = \left(1 - \hat{\gamma}_i\right)\hat{\sigma}_v^2,\ g_{2i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) = \left(\bar{\boldsymbol{X}}_i - \hat{\gamma}_i\bar{\boldsymbol{x}}_i\right)^T \left(\sum_{i=1}^{m} \boldsymbol{x}_i^T \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{x}_i\right)^{-1} \left(\bar{\boldsymbol{X}}_i - \hat{\gamma}_i\bar{\boldsymbol{x}}_i\right),$$

$$g_{3i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) = n_i^{-2}\left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 n_i^{-1}\right)^{-3} h\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right),$$

$$h\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) = \hat{\sigma}_e^4\,\mathrm{var}\left(\hat{\sigma}_v^2\right) - 2\hat{\sigma}_e^2\hat{\sigma}_v^2\,\mathrm{cov}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + \hat{\sigma}_v^4\,\mathrm{var}\left(\hat{\sigma}_e^2\right)$$

and

$$\hat{\boldsymbol{V}}_i = \hat{\sigma}_e^2 \boldsymbol{I}_{n_i} + \hat{\sigma}_v^2 \boldsymbol{1}_{n_i}\boldsymbol{1}_{n_i}^T.$$

The MSE estimator (2.9) is nearly unbiased (Rao 2003, Chapter 7) under non-informative sampling.

The MSE of the You-Rao (2002) pseudo-EBLUP estimator $\hat{\theta}_{iw}$ is estimated by

$$mse\left(\hat{\theta}_{iw}\right) = g_{1iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{2iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + 2g_{3iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) \tag{2.10}$$

where

$$g_{1iw}\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right)=\left(1-\hat{\gamma}_{iw}\right)\hat{\sigma}_v^2,\ g_{2iw}\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right)=\left(\bar{\boldsymbol{X}}_i-\hat{\gamma}_{iw}\bar{\boldsymbol{x}}_{iw}\right)^T\boldsymbol{\Phi}_w\left(\bar{\boldsymbol{X}}_i-\hat{\gamma}_{iw}\bar{\boldsymbol{x}}_{iw}\right),$$

$$\boldsymbol{\Phi}_w=\ \hat{\sigma}_e^2\left(\sum_{i=1}^{M}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}\boldsymbol{z}_{ij}^T\right)^{-1}\left(\sum_{i=1}^{M}\sum_{j=1}^{n_i}\boldsymbol{z}_{ij}\boldsymbol{z}_{ij}^T\right)\left\{\left(\sum_{i=1}^{M}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}\boldsymbol{z}_{ij}^T\right)^{-1}\right\}^T$$

$$+\hat{\sigma}_v^2\left(\sum_{i=1}^{M}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}\boldsymbol{z}_{ij}^T\right)^{-1}\left\{\sum_{i=1}^{M}\left(\sum_{j=1}^{n_i}\boldsymbol{z}_{ij}\right)\left(\sum_{j=1}^{n_i}\boldsymbol{z}_{ij}\right)^T\right\}\left\{\left(\sum_{i=1}^{M}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}\boldsymbol{z}_{ij}^T\right)^{-1}\right\}^T,$$

$$g_{3iw}\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right)=\hat{\gamma}_{iw}\left(1-\hat{\gamma}_{iw}\right)^2\hat{\sigma}_e^{-4}\hat{\sigma}_v^{-2}h\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right)\text{ and }\boldsymbol{z}_{ij}=w_{ij}\left(\boldsymbol{x}_{ij}-\hat{\gamma}_{iw}\bar{\boldsymbol{x}}_{iw}\right).$$

The estimator (2.10) ignores a covariance term in $\mathrm{MSE}\left(\hat{\theta}_{iw}\right)$. Torabi and Rao (2010) obtained an MSE estimator that accounts for the missing covariance term and that is nearly unbiased under non-informative sampling. However, it is computationally very intensive. It was not used here in the simulation study (Section 3) since it would have slowed down the simulations significantly. A few simulation trials, however, revealed the two MSE estimators give similar results under the simulation set-up used in Section 3.

3. SIMULATION STUDY

3.1 Implementation

A design-model (*pm*) approach was used for the simulation. That is, data are generated for the N population units according to a model, and a sample is then selected. The process of generating population data and selecting a sample is repeated R times. We next describe the steps to carry this out.

The population data of M=100 areas and N_i=100 units within area i were generated using the following simple nested error mean model:

$$y_{ij}=\mu+v_i+e_{ij},\ i=1,\mathrm{K}\ ,100,\ j=1,\mathrm{K}\ ,100 \tag{3.1}$$

where $\mu=0.5$, $v_i\overset{\text{iid}}{:}\ N\left(0,\sigma_v^2\right)$ and $e_{ij}\overset{\text{iid}}{:}\ N\left(0,\sigma_e^2\right)$ with $\sigma_v^2=0.5,\ \sigma_e^2=2$. A sample of size $n_i=5$ was selected within each area with probability proportional to size, using the Rao-Sampford πPS sampling scheme (Rao 1965, Sampford 1967). Following Asparouhov (2006), probabilities $\pi_{j|i}$ were generated according to two sampling mechanisms, denoted as invariant and non-invariant. A sampling mechanism is invariant across first-stage (clusters) units if the resulting sampling weight $w_{j|i}$ and the cluster random effects v_i are conditionally independent given the covariates $\boldsymbol{x}_{ij}$ (Asparouhov 2006, p. 444). If these conditions are not satisfied, then it is non-invariant. The selection probabilities within area i are given by

$$\pi_{j|i}=n_i\,b_{ij}\Big/\sum_{k=1}^{N_i}b_{ik}\ ,\ j=1,\mathrm{K}\ .N_i \tag{3.2}$$

where

$$b_{ij}=\left[1+\exp\left\{-\tau\left(\frac{1}{\alpha}e_{ij}+\sqrt{1-\frac{1}{\alpha^2}}e_{ij}^*\right)\right\}\right]^{-1} \tag{3.3}$$

for invariant selection, and

$$b_{ij}=\left[1+\exp\left\{-\tau\left(\frac{1}{\alpha}\left(v_i+e_{ij}\right)+\sqrt{1-\frac{1}{\alpha^2}}\left(v_i^*+e_{ij}^*\right)\right)\right\}\right]^{-1} \tag{3.4}$$

for non-invariant selection. The coefficient τ in (3.3) and (3.4), chosen as 0.5, ensured that the variation of the weights $w_{j|i}$ would not be too large within a simulation run. The error pair $\left(v_i^*, e_{ij}^*\right)$ was generated independently of $\left(v_i, e_{ij}\right)$, from the same distributions as v_i and e_{ij} to ensure that the weight variation would be comparable between various levels of α. If some of the $\pi_{j|i}$ exceeded one, they were set to one, and the probabilities were recomputed for the remaining units. The α-values, chosen as 1, 2, 3 or ∞, controlled the level of informativeness associated with the $\pi_{j|i}$'s. Increasing the α decreased informativeness, with $\alpha = \infty$ corresponding to non-informative sampling.

Two simple nested error models were fitted to the generated data for each selected sample. The first model is the mean model

$$y_{ij} = \mu + v_i + e_{ij}, \tag{3.5}$$

whereas the second model includes $z_{ij} = w_{j|i}$ as an auxiliary variable. This augmented model is given by

$$y_{ij} = \beta_0 + \beta_1 z_{ij} + v_i + e_{ij}. \tag{3.6}$$

For each model, the variance components σ_e^2 and σ_v^2 were estimated using the restricted maximum likelihood (REML) method. Simulation efficiency was improved by introducing various dependencies in the simulations. All four error components $\left(v_i, e_{ij}, v_i^*, e_{ij}^*\right)$ were first generated. Population *y*-values, as well as invariant and non-invariant probabilities of selection were then generated from those errors. For a given generated population, eight samples were selected: an invariant sample and a non-invariant sample for each value of α considered. For each of those samples, models (3.5) and (3.6) were fitted and the variance components σ_v^2 and σ_e^2 were estimated. Note that the weights $w_{j|i}$ obtained from (3.2) may not satisfy the condition (2.6), but we nonetheless fitted (2.6) in order to compute $\hat{b}$ needed in $\hat{\theta}_{i,ps}$ given by (2.8).

3.2 Results from the Empirical Study

Using the design-model (*pm*) approach, R =1000 samples were generated. From each sample r (r=1,...,R), the estimators $\hat{\theta}_{iu}^{(r)}$, $\hat{\theta}_{iw}^{(r)}$ and $\hat{\theta}_{ips}^{(r)}$ were computed for each small area i (i=1,...,M) under model (3.5). Only the first two estimators, $\hat{\theta}_{iu}^{(r)}$ and $\hat{\theta}_{iw}^{(r)}$, were computed under the augmented model (3.6) which includes the auxiliary variable z Performance of the estimators was judged using the average absolute bias ratio and the average MSE.

Average Absolute Bias Ratio. The average absolute bias ratio $\left(\overline{\text{ABR}}\right)$ was calculated as

$$\overline{ABR}\left(\hat{\theta}\right) = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{\left|\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_i^{(r)} - \overline{Y}_i^{(r)}\right)\right|}{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_i^{(r)} - \hat{\theta}_i^{(\bullet)}\right)^2}}\right)$$

where $\hat{\theta}_i^{(\bullet)} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_i^{(r)}$ and $\overline{Y}_i^{(r)}$ is the population mean associated with the i-th small area (i=1,...,M) and the r-th replicate (r=1,…, 1000). Values of the percent $\overline{\text{ABR}}$ are given in Table 1 which also includes the degree of informativeness of the selection mechanism as measured via an informativeness index. This measure was suggested by Asparouhov (2006) for estimating the degree of informativeness associated with the sampling design. It is given by $I_3\left(y\right) = \left|\mu - \hat{\mu}_u\right| \Big/ \sqrt{\hat{\sigma}_v^2 + \hat{\sigma}_e^2}$, where $\hat{\mu}_u = \hat{\beta}_u$ as defined in (2.3) under model (3.5). This informativeness measure is relatively independent of the cluster size and the sample size (Asparouhov, 2006).

The results are somewhat similar for invariant selection and non-invariant selection. Hence, results only for non-invariant selection are described. If the *z*-term is not included in the model, Table 1 shows that $\overline{\text{ABR}}$ associated with the EBLUP estimator is very large, 93%, when the design is very informative ($\alpha = 1$). The $\overline{\text{ABR}}$ of the EBLUP estimator gradually reduces to 3.3% as the design becomes non-informative ($\alpha = \infty$). On the other hand, for the pseudo-EBLUP estimator, $\overline{\text{ABR}}$ is much lower, 11%, when the design is very informative ($\alpha = 1$). This $\overline{\text{ABR}}$ reduces to 3.4 % as the design becomes non-informative ($\alpha = \infty$). The inclusion of the z-term in the model reduces $\overline{\text{ABR}}$ very significantly for the EBLUP estimator: $\overline{\text{ABR}}$ is 1.4% for the very informative design ($\alpha = 1$). Table 1 shows that it is better to use the EBLUP estimator under the augmented model (3.6) than to use the pseudo-EBLUP estimator under model (3.5) for a very informative design. On the other hand, for a given level of informativeness, EBLUP and pseudo-EBLUP estimators display similar $\overline{\text{ABR}}$ under the augmented model (3.6). The Pfeffermann and Sverchkov estimator is the best in terms of lower $\overline{\text{ABR}}$ over the whole spectrum of informativeness compared to the other two estimators based on model (3.5). However, it has somewhat larger $\overline{\text{ABR}}$ than the other two estimators if the latter are based on the augmented model (3.6) when the design is informative $(\alpha = 1, 2, 3)$.

Table 1. Average absolute bias ratio (%) of small area estimators $\left(\hat{\theta}_i\right)$

Sampling Scheme			**EBLUP** $\hat{\theta}_{iu}$		**Pseudo-EBLUP** $\hat{\theta}_{iw}$		**Pfeffermann and Sverchkov** $\hat{\theta}_{ips}$	**Informative Index** $I_3(y)$
Informative	α	**Selection**	**Without z**	**With z**	**Without z**	**With z**	**Without z**	
Yes	1	Invariant	81.3	2.1	11.1	1.9	5.7	29.5
		Non-invariant	92.9	1.4	10.5	1.6	7.6	30.0
	2	Invariant	40.4	2.9	5.6	2.9	5.5	14.2
		Non-invariant	42.2	2.7	5.6	2.9	5.9	14.3
	3	Invariant	27.7	3.1	4.5	3.1	4.2	9.6
		Non-invariant	28.3	3.0	4.7	3.1	4.1	9.7
No	∞	Invariant	3.1	3.1	3.2	3.2	3.1	0.2
		Non-invariant	3.3	3.3	3.4	3.5	3.3	0.1

Average Mean Squared Error. Values of average mean squared error $\left(\overline{\text{MSE}}\right)$ are reported in Table 2. $\overline{\text{MSE}}$ is calculated as

$$\overline{\text{MSE}}\left(\hat{\theta}\right) = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_i^{(r)} - \bar{Y}_i^{(r)}\right)^2\right)$$

We first discuss the results for the EBLUP estimator, $\hat{\theta}_{iu}$, based on the model (3.5). Table 2 shows that $\overline{\text{MSE}}$ of EBLUP is high (44%) under non-invariant selection when the design is very informative ($\alpha = 1$). $\overline{\text{MSE}}$ decreases as α increases, down to 23% under the non-informative design ($\alpha = \infty$). Results are similar under the invariant selection. Using the augmented model (3.6) lowers $\overline{\text{MSE}}$ for informative designs and this tendency is more pronounced with very informative designs as it was the case with $\overline{\text{ABR}}$.

The pseudo-EBLUP follows the same general trends, although its $\overline{\text{MSE}}$ does not reach the same extremes: a maximum of 26% when $\alpha = 1$ and model (3.5) is used and a minimum of 7.4% when $\alpha = 1$ and the augmented model (3.6) is used. Hence, under an informative design when model (3.5) is used, pseudo-EBLUP is preferred to EBLUP. If the augmented model (3.6) is used, then EBLUP is preferred. Table 2 shows that under non-informative sampling, EBLUP is always preferred. That was expected because EBLUP is by definition the "optimal" estimator.

Table 2. Average mean squared error (%) of small area estimators ($\hat{\theta}_i$)

Sampling Scheme			EBLUP $\hat{\theta}_{iu}$		Pseudo-EBLUP $\hat{\theta}_{iw}$		Pfeffermann and Sverchkov $\hat{\theta}_{ips}$
Informative	α	**Selection**	**Without z**	**With z**	**Without z**	**With z**	**Without z**
Yes	1	Invariant	41.1	6.4	25.4	7.4	21.6
		Non-invariant	44.2	6.2	26.2	7.8	24.9
	2	Invariant	27.0	19.7	23.9	20.7	22.4
		Non-invariant	28.2	19.0	24.4	20.3	23.5
	3	Invariant	24.8	21.5	23.9	22.5	22.7
		Non-invariant	25.2	21.1	24.1	22.2	23.1
No	∞	Invariant	22.6	22.7	23.7	23.7	22.7
		Non-invariant	22.8	23.0	24.0	24.1	23.0

The results associated with invariant selection are not very different from those obtained under non-invariant selection for EBLUP and pseudo-EBLUP. However, in the case of the estimator $\hat{\theta}_{ips}$, invariant selection results in better $\overline{\text{MSE}}$ than non-invariant selection. $\overline{\text{MSE}}$ does not vary much with α for this estimator, and it is better in terms of $\overline{\text{MSE}}$ than its competitors based on model (3.5), except when $\alpha = \infty$ in which case EBLUP is once more the best. However, EBLUP and pseudo-EBLUP based on the augmented model (3.6) perform better than the PS estimator when the design is informative.

Table 3. Average absolute relative bias (%) of the MSE estimators

Sampling Scheme			EBLUP $\hat{\theta}_{iu}$		Pseudo-EBLUP $\hat{\theta}_{iw}$	
Informative	α	**Selection**	**Without z**	**With z**	**Without z**	**With z**
Yes	1	Invariant	48.2	2.7	12.1	7.2
		Non-invariant	55.5	9.0	21.6	4.9
	2	Invariant	17.6	1.5	2.7	1.6
		Non-invariant	21.1	5.2	5.2	3.2
	3	Invariant	8.5	1.5	1.5	1.5
		Non-invariant	10.1	2.7	2.1	1.9
No	∞	Invariant	1.3	1.3	1.1	1.2
		Non-invariant	1.2	1.2	1.2	1.2

Average Absolute Relative Bias of Mean Squared Error Estimators. Two independent simulations were carried out to compare the average of the MSE estimators (2.9) and (2.10) with the empirical MSE. The average of the MSE estimators was based on $R_1 = 1,000$ samples, whereas the empirical MSE was based on $R_2 = 10,000$ samples. The average absolute relative bias $\left(\overline{\text{ARB}}\right)$, reported in Table 3, was computed over the small areas using

$$\overline{ARB}\left(mse\left(\hat{\theta}_i\right)\right) = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{\frac{1}{R_1}\sum_{r=1}^{R_1} mse\left(\hat{\theta}_i\right)^{(r)}}{\frac{1}{R_2}\sum_{r=1}^{R_2}\left(\hat{\theta}_i^{(r)} - \bar{Y}_i^{(r)}\right)^2} - 1\right|. \tag{3.9}$$

Table 3 shows that $\overline{\text{ARB}}$ of the MSE estimator associated with EBLUP under model (3.5) is very large when $\alpha = 1$, but $\overline{\text{ARB}}$ decreases as the design becomes less informative. For a given α, $\overline{\text{ARB}}$ is larger for non-invariant selection. The inclusion of the variable *z* greatly reduces $\overline{\text{ARB}}$. Similar observations hold for $\overline{\text{ARB}}$ of the MSE estimator associated with the pseudo-EBLUP estimator. However, its $\overline{\text{ARB}}$ is not as high as the one associated with the EBLUP estimator when the auxiliary data *z* is not included in the model. When the auxiliary variable *z* is included in the model, $\overline{\text{ARB}}$ associated with the pseudo-EBLUP is larger than the $\overline{\text{ARB}}$ associated with the EBLUP for the invariant selection; opposite holds for the non-invariant selection.

4. CONCLUDING REMARKS

In this paper, we first studied the bias and MSE of different small area estimators for various degrees of design informativeness under a nested error model for the population units. Estimators considered were the EBLUP, the pseudo-EBLUP (You & Rao, 2002) and an estimator given in Pfeffermann & Sverchkov (2007). The EBLUP and pseudo-EBLUP included or excluded the auxiliary data in the model. The sample design adapted the setting of Asparouhov (2006) to Rao-Sampford πPS sampling (Rao 1965 and Sampford 1967). Results from the simulations showed that design informativity can have a big impact on the bias and MSE of the EBLUP and the pseudo-EBLUP. Invariant selection and non-invariant selection did not produce very different results for the point estimates of the small area means. However, it can have a significant impact on the estimation of some of the parameters of the model as is shown in Asparouhov (2006) or Rao, Verret & Hidiroglou (2010). The simulations also showed that augmenting the model with the EBLUP and pseudo-EBLUP results in big gains over fitting the reduced version both in terms of bias and MSE. Pfeffermann & Sverchkov's point estimator gave good results in terms of bias compared to the other estimators.

We then evaluated the MSE estimators of the EBLUP and the pseudo-EBLUP. If the model (3.5) is fitted, it is better to use the pseudo-EBLUP to dampen the impact of design informativeness on the average absolute relative bias of the estimated MSE. On the other hand, if the augmented model (3.6) is fitted, the impact of design informativeness is reduced significantly for both estimators. Nevertheless, the pseudo-EBLUP has the desirable property of design-consistency as the area sample size increases, unlike the EBLUP (You & Rao, 2002).

All estimators considered have some flaws. The estimated MSE associated with the EBLUP, based on the model (3.5) without the auxiliary variable *z* is significantly biased if the sampling design is informative. In the case of the pseudo-EBLUP, the relative bias of the estimated MSE is much smaller under informative sampling, except under non-invariant selection and α=1. The form of the Pfeffermann & Sverchkov estimator depends on the assumed relationship between the sampling weight $w_{j|i}$ and the variables y_{ij} and x_{ij} of the unit level model (2.1).

REFERENCES

Asparouhov, T. (2006). General multi-level modelling with sampling weights. *Communication in Statistics, Theory and Methods*, 439-460.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An Error Component Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, **83**, 28–36.

Pfeffermann D. and Sverchkov M. (2007). Small-Area Estimation under Informative Probability Sampling of Areas and within the Selected Areas. *Journal of the American Statistical Association*, Vol. 102, No. 480, 1427-1439.

Rao, J.N.K. (1965). On Two Simple Schemes of Unequal Probability Sampling without Replacement. *Journal of the Indian Statistical Association* 3, 173-180.

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

Rao, J.N.K., Verret, F. & Hidiroglou, M.A. (2010). A weighted estimating equations approach to inference for two-level models from survey data. *Proceedings of the survey method section, SSC annual meeting, 2010*.

Sampford, M.R. (1967). On Sampling without Replacement with Unequal Probabilities of Selection. *Biometrika*, **54**, 499-513.

Torabi, M. and Rao, J.N.K. (2010). Mean squared error estimators of small area means using survey weights. Scheduled to appear in the December 2010 issue of the *Canadian Journal of Statistics*.

You Y. and Rao, J.N.K. (2002). A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights. *The Canadian Journal of Statistics*, Vol. 30, No. 3, 431-439.