STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 17, No. 1, pp. 133–154

A COMPARISON OF SMALL AREA AND CALIBRATION ESTIMATORS VIA SIMULATION

M. A. Hidiroglou[1], V. M. Estevao [2]

ABSTRACT

Domain estimates are typically obtained using calibration estimators that are direct or modified direct. They are direct if they strictly use data within the domain of interest. They are modified direct if they use both data within and outside the domain of interest. An alternative way of producing these estimates is through small area procedures. In this article, we compare the performance of these two approaches via a simulation. The population is generated using a hierarchical model that includes both area effects and unit level random errors. The population is made up of mutually exclusive domains of different sizes, ranging from a small number of units to a large number of units. We select many independent simple random samples of fixed size from the population and compute various estimates for each sample using the available auxiliary information. The estimates computed for the simulation included the Horvitz-Thompson estimator, the synthetic estimator (indirect estimate), calibration estimators, and unit level based estimators (small area estimate). The performance of these estimators is summarized based on their design- based properties.

Key words: area level, unit level, calibration estimates, small area estimates, simulation.

1. Introduction

Domain estimates at Statistics Canada are typically obtained using well-established methods based on calibration estimation. The calibration is direct or modified direct. It is direct if it is based on data within the domain of interest. It is modified direct if it is based on data within and outside the domain of interest. These methods can be viewed as design-based procedures as the variance of the resulting estimators is evaluated under the randomization distribution. The

[1] Michael A. Hidiroglou, Statistical Research and Innovation Division, Statistics Canada, 16 D, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: hidirog@yahoo.ca.

[2] Victor M. Estevao, Statistical Research and Innovation Division, Statistics Canada, 16 D, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: Victor.Estevao@statcan.gc.ca.

randomization distribution of an estimator is the distribution over all possible samples that could be selected from the target population of interest under the sampling design used to select the sample, with the population parameters considered as fixed values. Another way of producing these estimates is through small area methods. These methods are particularly important when the sample size in the domains is "small." They can improve the reliability of the direct estimates provided that the variable of interest is well correlated with auxiliary variables $\boldsymbol{x}$ that are available from administrative or other files. Small area estimation essentially combines direct estimates with model-based estimates in an optimal manner.

The model-based estimates involve known population totals (auxiliary data) and estimates of the regression between the variable of interest and the auxiliary data across the small areas. In general, these models are classified into two groups: unit level models and area level models. Unit level models are generally based on observation units (e.g., persons or companies) from the survey and auxiliary variables associated with each observation, whereas area level models are based on direct survey estimates aggregated from the unit level data and related area level auxiliary variables; see Rao (2003) for an overview of small area models. The more recent literature that covers empirical assessment of the properties of various small area and domain estimators includes Lehtonen and Veijanen (2009), Datta (2009), and Pfeffermann (2013). Lehtonen and Veijanen (2009) focused on design-based methods (calibration and regression) using auxiliary data. They reviewed the work on the extension of the linear form of the Generalized Regression Estimator (GREG) given in Särndal et al. (1992) to include logistic, multinomial logistic and mixed models for domain estimation. Datta (2009) reviewed the development of model-based procedures to obtain small area estimates. Datta focussed in particular on the theoretical properties of the resulting estimators. Pfeffermann (2013) reviewed both design-based and model-based procedures, as well as recent developments in these two procedures.

Domain estimates are currently obtained via design-based procedures at Statistics Canada. However, the increasing requirement for producing estimates for "small domains" has encouraged the need to adopt model-based procedures. A SAS-based prototype (Estevao et al. 2014) has been recently developed at Statistics Canada to respond to these requirements. The prototype currently incorporates two well-known methods initially developed by Fay and Herriot (1979) for area level estimation, and Battese, Harter, and Fuller (1988) for unit level estimation. Although the theoretical properties of the estimators included in the prototype are known, they were investigated via a simulation. In the simulations, we looked at the properties of estimators of domain totals. We compared model-based small area estimators with traditional estimators through simulation. The latter included the Horvitz-Thompson estimator, two calibration estimators, the modified regression estimator and the synthetic estimator. The small area estimators are the EBLUP and Pseudo EBLUP estimators based on a unit level model. More details on all of these estimators are given in section 2.

The simulation setup and results are reported in section 3. Section 4 provides a few conclusions from our findings.

2. Sample design

Large scale surveys are designed to satisfy reliability requirements for some subsets (domains) of the population. Examples of these subsets include partitions below the level of the initial geographical / industrial detail requested by the client. If such subsets are required before the sample is selected, then such domains are labelled as *planned domains* (Singh, Gambino and Mantel 1994). Such planned domains will have some of the sample allocated to them to obtain unbiased estimates with the required precision using direct estimation procedures. If these domains are identified after the sample has been selected, they will be known as *unplanned domains*. Note that, in any event, unplanned domains will exist for most surveys. An example taken from business surveys is a change of industry during data collection. A business initially classified as industry A becomes industry B. Such a business would be tabulated as part of the businesses of type B, but would retain its original sampling weight. Another example, taken from household surveys, would be the arbitrary production of estimates below a geographical level that was not part of the allocation process of the sample. Traditional or small area estimators can be used for either planned or unplanned domains.

As domain estimation for most surveys at Statistics Canada is mostly of the unplanned type, we have designed our simulation to reflect this tendency: that is no units are allocated to them prior to sample selection. Domain estimates are produced after sample selection, and the number of sampled units falling in each domain is a random variable. Our simulation reflects this point, and we used the simplest sample design to carry it out. We drew repeated samples *s* of size *n* from the population *U* of size *N* using simple random sampling without replacement. The weight associated with unit $j \in U$ is denoted as w_j. Let s_d, $d = 1, 2, \ldots, D$, be the portion of the sample *s* that overlaps with domain U_d (of known size N_d). Let the realized sample size in domain U_d be n_d. The survey design weight associated with a unit $j \in U_d$ is w_j. The data in the population are denoted as $\left(y_j, \boldsymbol{x}_j\right)$ for each element $j \in U$. The *y* variable is the one of interest, while $\boldsymbol{x}$ is the vector of auxiliary data. Computation of domain statistics can be obtained using the operators (i.e.: mean and variance) in regular estimation via the following transformation. In domain U_d, we denote the variable of interest as y_{dj} where $y_{dj} = y_j$ if $j \in U_d$ and 0 otherwise. The associated vector of auxiliary variables is defined as $\boldsymbol{x}_{dj}$ where $\boldsymbol{x}_{dj} = \boldsymbol{x}_j$ if $j \in U_d$ and **0** otherwise.

The objective of the present study is to compare the properties of model-based small area estimators for domains with those traditionally used in survey estimation. We considered seven estimators of the domain total $Y_d = \sum_{j \in U} y_{dj}$: four are traditional estimators and three are small area estimators. We first present the traditional estimators.

2.1. Traditional estimators

Horvitz-Thompson: The Horvitz-Thompson estimator $\hat{Y}_{d\,HT}, d = 1, 2, ..., D$, uses no auxiliary information. It is defined as $\hat{Y}_{d\,HT} = \sum\limits_{j \in s} w_j y_{dj}$ if $n_d > 0$ and 0 otherwise. We set $\hat{Y}_{d\,HT}$ to 0 if there are no sampled units in the domain, ensuring unbiased estimation over all samples *s* drawn from *U*. Although this estimator is unbiased, it produces inefficient estimates.

Calibration Estimators: We consider two calibration estimators, $\hat{Y}_{d\,CAL\,U_d}$ and $\hat{Y}_{d\,CAL\,U}$, that use auxiliary information at different levels. They are applications of calibration given in Deville and Särndal (1992) adapted to domain estimation. The direct estimator $\hat{Y}_{d\,CAL\,U_d}$ uses auxiliary information at the domain level, while the modified direct estimator $\hat{Y}_{d\,CAL\,U}$ uses information at the population level. Estimator $\hat{Y}_{d\,CAL\,U_d}$ is known to be more efficient than $\hat{Y}_{d\,CAL\,U}$. However, estimator $\hat{Y}_{d\,CAL\,U_d}$ has some drawbacks. It is not always possible to obtain auxiliary information at the domain level. Even if this information is available, we cannot produce estimates using $\hat{Y}_{d\,CAL\,U_d}$ if there are no sample units in the domain. Furthermore, this estimator can produce erratic values when there are only a few units in the domain. To prevent this, we need to make sure that the number of units in the domain is larger than the number of auxiliary variables. As a minimal requirement, given that there are two auxiliary variables (intercept, *x*), $\hat{Y}_{d\,CAL\,U_d}$ can be estimated only if there are 3 or more units in a domain. Otherwise, we cannot produce a value, and we set it to missing. This means that we only work with a subset of all possible samples. If we set the value of $\hat{Y}_{d\,CAL\,U_d}$ to 0 when there is an insufficient number of observations n_d in domain U_d, this would result in a biased estimator. As for $\hat{Y}_{d\,CAL\,U}$, when there are no sample units in the domain, we set the value of this estimator to 0. This ensures that it is approximately design unbiased for the domain total. Estimator $\hat{Y}_{dCALU_d}$, $d=1,2,...,D$, is given by:

$$\hat{Y}_{dCALU_d}=\begin{cases}\sum\limits_{j\in s} w_j y_{dj}+(\boldsymbol{X}_d-\hat{\boldsymbol{X}}_{dHT})^T\hat{\boldsymbol{\beta}}_{dCALU_d} & \text{if } n_d\geq 3\\ .\ \text{(missing)} & \text{if } n_d<3\end{cases}$$

with $\boldsymbol{X}_d=\sum\limits_{j\in U}\boldsymbol{x}_{dj}$, $\hat{\boldsymbol{X}}_{dHT}=\sum\limits_{j\in s} w_j\boldsymbol{x}_{dj}$ and

$$\hat{\boldsymbol{\beta}}_{dCALU_d}=\left(\sum_{j\in s}\frac{w_j\boldsymbol{x}_{dj}\boldsymbol{x}_{dj}^T}{c_j}\right)^{-1}\sum_{j\in s}\frac{w_j\boldsymbol{x}_{dj}y_{dj}}{c_j}.$$

Estimator $\hat{Y}_{dCALU}$, $d=1,2,...,D$, is given by:

$$\hat{Y}_{dCALU}=\begin{cases}\sum\limits_{j\in s} w_j y_{dj}+(\boldsymbol{X}-\hat{\boldsymbol{X}}_{HT})^T\hat{\boldsymbol{\beta}}_{CALU} & \text{if } n_d>0\\ 0 & \text{if } n_d=0\end{cases}$$

with $\boldsymbol{X}=\sum\limits_{j\in U}\boldsymbol{x}_j$, $\hat{\boldsymbol{X}}_{HT}=\sum\limits_{j\in s} w_j\boldsymbol{x}_j$ and $\hat{\boldsymbol{\beta}}_{CALU}=\left(\sum\limits_{j\in s}\frac{w_j\boldsymbol{x}_j\boldsymbol{x}_j^T}{c_j}\right)^{-1}\sum\limits_{j\in s}\frac{w_j\boldsymbol{x}_j y_j}{c_j}$.

Modified Regression (REG): The modified regression estimator $\hat{Y}_{dREG}$, $d=1,2,...,D$, is due to Woodruff (1966). It is of interest as it was used to produce area breakdowns of the monthly national estimates of US Census Bureau retail trade survey. Note that it is a modified direct estimator. Singh and Mian (2003) points out that it can be viewed as a calibration estimator $\sum_s\tilde{w}_{dj}y_j$, where the calibration weight $\tilde{w}_{dj}$ is obtained by minimizing the chi-squared distance $\sum_s c_j\left(w_j a_{dj}-\tilde{w}_{dj}\right)/w_j$, subject to the constraints $\sum_s\tilde{w}_{dj}\boldsymbol{x}_j=\boldsymbol{X}_d$: here a_{dj} is the domain indicator variable. Estimator $\hat{Y}_{dREG}$ is design-unbiased as the overall sample size increases. It is given by:

$$\hat{Y}_{dREG}=\begin{cases}\sum\limits_{j\in s_d} w_{dj}y_{dj}+(\boldsymbol{X}_d-\hat{\boldsymbol{X}}_{dHT})^T\hat{\boldsymbol{\beta}}_{REG} & \text{if } n_d>0\\ \boldsymbol{X}_d^T\hat{\boldsymbol{\beta}}_{REG} & \text{if } n_d=0\end{cases}$$

with $\boldsymbol{X}_d = \sum_{j\in U_d} \boldsymbol{x}_{dj}$, $\hat{\boldsymbol{X}}_{dHT} = \sum_{j\in s} w_j \boldsymbol{x}_{dj}$ and $\hat{\boldsymbol{\beta}}_{REG} = \left(\sum_{j\in s} \frac{w_j \boldsymbol{x}_j \boldsymbol{x}_j^T}{c_j} \right)^{-1} \sum_{j\in s} \frac{w_j \boldsymbol{x}_j y_j}{c_j}$.

The c_j term, $c_j > 0$, associated with the estimators that use auxiliary data reflects that the error terms e_j in the implied working model are distributed independently with mean zero and variance $c_j^2 \sigma_e^2$.

2.2. Small area estimators

The simplest small area estimator is the synthetic estimator (SYN), $\hat{Y}_{dSYN}$, $d = 1, 2, ..., D.$ It is given by $\hat{Y}_{dSYN} = \boldsymbol{X}_d^T \hat{\boldsymbol{\beta}}_{SYN}$ where $\boldsymbol{X}_d = \sum_{j\in U_d} \boldsymbol{x}_{dj}$ and $\hat{\boldsymbol{\beta}}_{SYN} = \left(\sum_{j\in s} \frac{w_j \boldsymbol{x}_j \boldsymbol{x}_j^T}{c_j} \right)^{-1} \sum_{j\in s} \frac{w_j \boldsymbol{x}_j y_j}{c_j}$. This estimator is design-biased, given by $Bias\ (\hat{Y}_{dSYN}) \square \boldsymbol{X}_d^T \boldsymbol{B} - Y_d$, where $\boldsymbol{B} = \left(\sum_{j\in U} \frac{\boldsymbol{x}_j \boldsymbol{x}_j^T}{c_j} \right)^{-1} \sum_{j\in U} \frac{\boldsymbol{x}_j y_j}{c_j}$ is the population regression vector.

The next two small area estimators are based on a hierarchical model given by:

$$y_{dj} = \boldsymbol{x}_{dj}^T \boldsymbol{\beta} + v_d + e_{dj} \, , \tag{1}$$

where $v_d \overset{iid}{\square} N(0, \sigma_v^2), e_{dj} \overset{iid}{\square} N(0, c_{dj}^2 \sigma_e^2)$, and c_{dj} accounts for possible heterogeneity of the e_{dj} residuals.

In our application of this model, the areas are our domains of interest. The quantity $\boldsymbol{x}_{dj}^T \boldsymbol{\beta}$ is the fixed effect which is assumed to be a linear combination of the auxiliary variables $\boldsymbol{x}_{ij}$. The residuals v_d and e_{dj} are respectively the random effect for the area *d* and the random errors for unit *j* in area *d*. The term c_{dj}^2 translates to $a_{dj} = c_{dj}^{-2}$ in the various formulas that follow.

Empirical Best Linear Unbiased Predictor (EBLUP): This estimator denoted as $\hat{Y}_{d\,EBLUP}$, $d = 1,2,...,D$, is given in Rao (2003, p.136). It is an extension of the Battese, Harter, and Fuller (1988) estimator when the error structure of the residuals is not homogeneous. It is given by:

$$\hat{Y}_{d\,EBLUP} = \begin{cases} N_d\,\{\bar{\boldsymbol{X}}_d^T\,\hat{\boldsymbol{\beta}}_{EBLUP} + \hat{\gamma}_{da}(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T\,\hat{\boldsymbol{\beta}}_{EBLUP})\} & \text{if } n_d > 0 \\ \boldsymbol{X}_d^T\,\hat{\boldsymbol{\beta}}_{EBLUP} & \text{if } n_d = 0 \end{cases}$$

The terms making up $\hat{Y}_{d\,EBLUP}$ include N_d , $\boldsymbol{X}_d$, $\hat{\gamma}_{da}$ $\bar{y}_{da}$ $\bar{\boldsymbol{x}}_{da}$, and $\hat{\boldsymbol{\beta}}_{EBLUP}$. These terms are defined as follows: $\bar{y}_{da} = \dfrac{\sum_{j\in s_d} a_{dj} y_{dj}}{\sum_{j\in s_d} a_{dj}}$, $\bar{\boldsymbol{x}}_{da} = \dfrac{\sum_{j\in s_d} a_{dj} \boldsymbol{x}_{dj}}{\sum_{j\in s_d} a_{dj}}$, and $\hat{\gamma}_{da} = \dfrac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2\,\delta_{da}^2}$ where $\delta_{da}^2 = \dfrac{1}{\sum_{j\in s_d} a_{dj}}$. The estimated regression vector is given by:

$$\hat{\boldsymbol{\beta}}_{EBLUP} = \left(\sum_{d=1}^{D} \sum_{j\in s_d} a_{dj}\,(\boldsymbol{x}_{dj} - \hat{\gamma}_{da}\,\bar{\boldsymbol{x}}_{da})\,\boldsymbol{x}_{dj}^T \right)^{-1} \sum_{d=1}^{D} \sum_{j\in s_d} a_{dj}\,(\boldsymbol{x}_{dj} - \hat{\gamma}_{da}\,\bar{\boldsymbol{x}}_{da})\,y_{dj} \ .$$

This estimator is not design consistent, unless the sampling design is self-weighting.

Pseudo-EBLUP (PEBLUP): This estimator denoted as $\hat{Y}_{d\,PEBLUP}$, $d = 1,2,...,D$, is an extension of the Pseudo-EBLUP estimator given in You and Rao (2002). It accounts for the heterogeneity of the e_{dj} residuals in model (1). It includes the survey weights w_j , $j \in s$, in the regression coefficient and the parameter estimate.

$$\hat{Y}_{d\,PEBLUP} = \begin{cases} N_d\,\{\bar{\boldsymbol{X}}_d^T\,\hat{\boldsymbol{\beta}}_{PEBLUP} + \hat{\gamma}_{dw}(\bar{y}_{dw} - \bar{\boldsymbol{x}}_{dw}^T\,\bar{y}_{dw})\} & \text{if } n_d > 0 \\ \boldsymbol{X}_d^T\,\hat{\boldsymbol{\beta}}_{PEBLUP} & \text{if } n_d = 0 \end{cases}$$

The terms making up $\hat{Y}_{d\,PEBLUP}$ include N_d , $\boldsymbol{X}_d$, $\bar{y}_{dw}$, $\bar{\boldsymbol{x}}_{dw}$, $\bar{y}_{dw}$, and $\hat{\boldsymbol{\beta}}_{PEBLUP}$.

These terms are defined as follows: $\bar{y}_{dw} = \dfrac{\sum_{j\in s_d} w_j y_{dj}}{\sum_{j\in s_d} w_j}$, $\bar{\boldsymbol{x}}_{dw} = \dfrac{\sum_{j\in s_d} w_j \boldsymbol{x}_{dj}}{\sum_{j\in s_d} w_j}$,

$\hat{\gamma}_{dw} = \dfrac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2\,\delta_{dw}^2}$, where $\delta_{dw}^2 = \dfrac{\sum_{j\in s_d} w_j^2 \big/ a_{dj}}{\left(\sum_{j\in s_d} w_j\right)^2}$ for $d = 1,2,...,D$.

The estimated regression vector is given by:

$$\hat{\boldsymbol{\beta}}_{PEBLUP} = \left(\sum_{d=1}^{D} \sum_{j \in s_d} w_j a_{dj} (\boldsymbol{x}_{dj} - \hat{\gamma}_{dwa} \bar{\boldsymbol{x}}_{dwa}) \boldsymbol{x}_{ij}^T \right)^{-1} \sum_{d=1}^{D} \sum_{j \in s_d} w_j a_{dj} (\boldsymbol{x}_{dj} - \hat{\gamma}_{dwa} \bar{\boldsymbol{x}}_{dwa}) y_{dj}$$

with $\bar{\boldsymbol{x}}_{dwa} = \dfrac{\sum_{j \in s_d} w_j a_{dj} \boldsymbol{x}_{dj}}{\sum_{j \in s_d} w_j a_{dj}}$ and $\hat{\gamma}_{dwa} = \dfrac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \delta_{dwa}^2}$

where $\delta_{dwa}^2 = \dfrac{\sum_{j \in s_d} \dfrac{(w_j a_{dj})^2}{a_{dj}}}{\left(\sum_{j \in s_d} w_j a_{dj} \right)^2}$.

This estimator is design consistent.

3. Simulation

Surveys produced at Statistics Canada can be as simple as stratified one stage simple random sampling designs typically used for business surveys to the more complex stratified multi-stage design with unequal selection probabilities at each stage typically used for household surveys. We opted for a single stage simple random sample selected from the population, as it is a simplification of the sample designs used for business surveys. Had we chosen a sampling design with unequal weights, we would have had to account for the possible impact of informative sampling on the small area estimators using the procedure given in Pfeffermann and Sverchkov (2007). Verret, Rao, and Hidiroglou (2015) used a simpler procedure than the one given in Pfeffermann and Sverchkov (2007). Their procedure accounted for unequal selection probabilities for model-based small area estimators by incorporating them into the model. Their simulation used a design-model (*pm*) approach. Their results showed that incorporating the unequal selection probabilities significantly improved the performance (average absolute bias and average RMSE) of EBLUP, but had marginal impact on PEBLUP.

3.1 Population Generation and Sample Selection

A population *U* consisting of 4,640 units was created by generating data (x_{ij}, y_{ij}) for three separate subsets of the population (groups) with different intercepts and slopes. Each group was split into mutually exclusive and exhaustive domains as follows: Group 1 was split into nine domains $U_1,...,U_9$; Group 2 was split into ten domains $U_{10},...,U_{19}$; and Group 3 was split into ten domains $U_{20},...,U_{29}$. The three groups resulted in a total of *D*=29 domains that were mutually exclusive and exhaustive. The number of units in each domain, N_d, was allocated in a monotonic manner: domain U_1 had 20 units; domain U_2 had 30 units; and domain U_{29} had 300 units. In our simulation the auxiliary data $\mathbf{x}$ consisted of two auxiliary variables. The first one had the fixed value of one to represent the intercept in the model. The second one, x , represented the available auxiliary data in the population. The auxiliary variable x in each group was generated from a $Gamma\,(\alpha = 5, \beta = 10)$ distribution with mean $\alpha\beta = 50$ and variance $\alpha\beta^2 = 500$. The variable of interest y was generated using the model

$$y_{dj} = \beta_{0,\ell} + \beta_{1,\ell}\, x_{dj} + v_d + e_{dj} :\ \ell=1,2,3 \tag{2}$$

where $v_d \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{dj} \overset{iid}{\sim} N(0, c_{dj}^2\, \sigma_e^2)$.

We used $\sigma_v^2 = \sigma_e^2 = 20^2 = 400$ and set c_{dj}^2 equal to x_{dj}. The following table summarizes how the population was split into the three groups of domains.

Table 1. Groups, associated domains and regression parameters

***Group* (ℓ)**	***Domains in Group***	$\beta_{0,\ell}$	$\beta_{1,\ell}$
1	U_d for d= 1,..., 9	200	30
2	U_d for d=10,...,19	300	20
3	U_d for d =20,...,29	400	10

A plot of the generated population is shown in Figure 1. The units in the groups are shown respectively in green, blue and yellow. The three regression lines are shown in red. Without the colours to identify the groups, one might be inclined to think that the population was generated under a model with a single auxiliary variable (one intercept and slope) as shown in the inset.

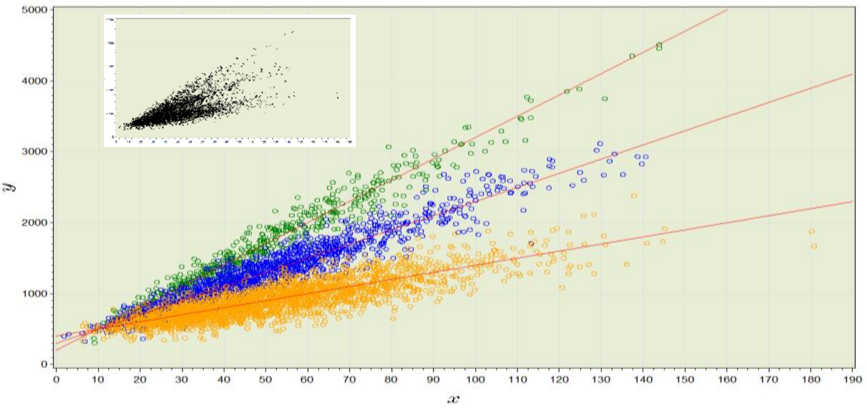

Figure 1. Plot of y vs. *x* for population in the simulation study

We ran two separate simulation "runs" to reflect that two possible models could be fitted for the selected samples. We denote these simulations as runs 1 and 2. In the first run (simulation run 1), we assumed that the model could be fitted using $\boldsymbol{x}_{dj} = (1, x_{dj})$ as auxiliary data; this is not correct as the population was generated on the basis of three different regressions. In the second run (simulation run 2), we acknowledged that there were three separate models and used a set of auxiliary variables reflecting the manner in which the population values were generated; this fit is correct. This meant using a set of dummy-coded auxiliary variables defined as follows for each unit:

$$\boldsymbol{x}_{dj}^{T} = \begin{cases} \left(1, 0, 0, x_{dj}, 0, 0\right) & \text{if } j \in U_d \in \text{group 1} \\ \left(0, 1, 0, 0, x_{dj}, 0\right) & \text{if } j \in U_d \in \text{group 2} \\ \left(0, 0, 1, 0, 0, x_{dj}\right) & \text{if } j \in U_d \in \text{group 3} \end{cases} \tag{3}$$

In the small area estimation model given by equation (1), the use of this $\boldsymbol{x}_{dj}$ implies the following regression coefficient $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6\right)^T$ for the fixed effects. For the synthetic estimator and the calibration estimators, we set $c_{ij} = x_{ij}$ to reflect the heterogeneity of the model errors.

Each simulation run involved the selection of *R*=100,000 independent samples and the computation of various estimates for each sample. Each sample was a simple random sample *s* of size *n* selected without replacement from *U*. We used sample sizes *n*=232 (5%), *n*=464 (10%), *n*=696 (15%) and *n*=928 (20%), where the sampling fractions are indicated in brackets. These are within the range of the sampling fractions typically used by business surveys.

The sample units in domain U_d are denoted by s_d with $s = \bigcup_{d=1}^{D} s_d$. We observed n_d units in U_d where $0 \le n_d \le N_d$ and $n = \sum_{d=1}^{D} n_d$. Under simple random sampling without replacement, the n_d follow a multivariate hypergeometric distribution with probability mass function $\prod_{d=1}^{D} \binom{N_d}{n_d} \Big/ \binom{N}{n}$.

The following table shows the probability of observing $n_d = 0$, $n_d = 1$ or $n_d = 2$ in the three smallest domains when the sample size *n* is 232.

Table 4. Probabilities in the 3 smallest domains when *n* = 232

Probability	U_1 with $N_1 = 20$	U_2 with $N_2 = 30$	U_3 with $N_3 = 40$
Prob $(n_d = 0)$	0.358	0.214	0.127
Prob $(n_d = 1)$	0.378	0.339	0.271
Prob $(n_d = 2)$	0.189	0.260	0.279

Using table 4, for the smallest domain U_1, $\hat{Y}_{d\,HT}$ and $\hat{Y}_{d\,CAL\,U}$ would be equal to zero about 36% of the time. Note that this probability decreases rapidly as the domain population size N_d increases. Since we require $n_d \geq 3$, we cannot produce an estimate for $\hat{Y}_{d\,CAL\,U_d}$ in approximately 92.5% of the samples selected in the smallest domain U_1. This probability decreases rapidly as the domain population size N_d increases.

3.2. Simulation statistics

For each selected sample in each simulation run *r* = 1,...,*R* (*R*=100,000), we computed estimates of Y_d for the seven estimators. Denote $\hat{Y}^{(r)}_{d\,EST}$ as the estimate produced for the r^{th} sample, $r = 1, 2, \ldots R$, where the subscript '*EST*' is a placeholder for any one of the seven estimators. For each domain *d*=1,...,29, we computed the bias as:

$$Bias(\hat{Y}_{d\,EST}) = R^{-1}\sum_{r=1}^{R}\hat{Y}^{(r)}_{d\,EST} - Y_d$$

and the mean squared error as

$$MSE(\hat{Y}_{d\,EST}) = R^{-1}\left(\sum_{r=1}^{R}\hat{Y}^{(r)}_{d\,EST} - Y_d\right)^2 .$$

For each estimator, $\hat{Y}_{d\,EST}$, we also computed the following summary statistics across all domains and simulated samples. These were the average absolute relative bias, the average coefficient of variation and the average relative efficiency denoted as $\overline{ARB}(\hat{Y}_{EST})$, $\overline{CV}(\hat{Y}_{EST})$ and $\overline{RE}(\hat{Y}_{EST})$ respectively.

These were computed as follows:

$$\overline{ARB}(\hat{Y}_{EST}) = \frac{1}{D}\sum_{d=1}^{D} ARB(\hat{Y}_{d\,EST}) \quad \text{where } ARB(\hat{Y}_{d\,EST}) = \left| \frac{Bias(\hat{Y}_{d\,EST})}{Y_d} \right|$$

$$\overline{CV}(\hat{Y}_{EST}) = \frac{1}{D}\sum_{d=1}^{D} CV(\hat{Y}_{d\,EST}) \quad \text{where } CV(\hat{Y}_{d\,EST}) = \frac{\sqrt{MSE(\hat{Y}_{d\,EST})}}{Y_d}$$

$$\overline{RE}(\hat{Y}_{EST}) = \sqrt{\frac{\overline{MSE}(\hat{Y}_{HT})}{\overline{MSE}(\hat{Y}_{EST})}} \quad \text{where } \overline{MSE}(\hat{Y}_{EST}) = \frac{1}{D}\sum_{d=1}^{D} MSE(\hat{Y}_{d\,EST}) \tag{4}$$

The statistic $\overline{RE}(\hat{Y}_{EST})$ measures the average efficiency of each estimator relative to the Horvitz-Thompson estimator. Since $\hat{Y}_{d\,HT}$ is known to have the least efficiency among these seven estimators, this measure is a number larger than or equal to 1.

3.3. Simulation Results

Tables 5, 6 and 7 show the differences between the two runs using the summary statistics described in the previous section. The results are discussed after each of these three tables for runs 1 and 2.

Table 5. Average Absolute Relative Bias $\overline{ARB}(\hat{Y}_{dEST})$

		Traditional Domain Estimators				**Small Area Estimators**		
Sample Size	**Run**	$\hat{Y}_{dHT}$	$\hat{Y}_{d\,CALU_d}$	$\hat{Y}_{d\,CALU}$	$\hat{Y}_{d\,REG}$	$\hat{Y}_{dSYN}$	$\hat{Y}_{dEBLUP}$	$\hat{Y}_{dPEBLUP}$
232	1	0.12	0.16	0.15	0.19	24.18	7.58	4.12
	2	0.11	0.15	0.36	0.05	1.33	1.07	1.08
464	1	0.08	0.08	0.09	0.10	24.18	6.71	2.24
	2	0.06	0.07	0.19	0.02	1.33	0.95	0.96
696	1	0.06	0.05	0.06	0.06	24.18	6.43	1.52
	2	0.05	0.04	0.11	0.02	1.33	0.84	0.86
928	1	0.06	0.03	0.06	0.04	24.18	6.29	1.14
	2	0.05	0.03	0.09	0.01	1.33	0.76	0.77

Model does not fit (**Run 1**): The small area estimators $\hat{Y}_{dSYN}$, $\hat{Y}_{dEBLUP}$, and $\hat{Y}_{dPEBLUP}$ have the largest $\overline{ARB}$ s. In particular, $\hat{Y}_{dSYN}$ has the highest $\overline{ARB}$. The $\overline{ARB}$ decreases as the sample size increases for $\hat{Y}_{dEBLUP}$ and $\hat{Y}_{dPEBLUP}$, whereas it remains constant (as expected) for $\hat{Y}_{dSYN}$. The $\overline{ARB}$ associated with $\hat{Y}_{dPEBLUP}$ decreases more rapidly than the one associated with $\hat{Y}_{dEBLUP}$ as the sample size increases. The $\overline{ARB}$ associated with the traditional domain estimators is quite small: it also decreases as the sample size increases.

Model fits (**Run 2**): The $\overline{ARB}$ s associated with the small area estimators have significantly decreased. However, they are still higher than those associated with the traditional domain estimators. $\hat{Y}_{d\,REG}$ has the smallest $\overline{ARB}$ amongst all the estimators. As noted in run 1, the $\overline{ARB}$ decreases as the sample size increases for all the estimators.

Table 6. Average Coefficient of Variation $\overline{CV}(\hat{Y}_{dEST})$

		Traditional Domain Estimators			**Small Area Estimators**			
Sample Size	**Run**	$\hat{Y}_{dHT}$	$\hat{Y}_{d\,CALU_d}$	$\hat{Y}_{d\,CAL\,U}$	$\hat{Y}_{d\,REG}$	$\hat{Y}_{dSYN}$	$\hat{Y}_{dEBLUP}$	$\hat{Y}_{dPEBLUP}$
232	1	42.79	6.57	42.81	12.82	24.27	9.90	7.93
	2	42.77	6.39	42.04	4.47	2.17	2.22	2.21
464	1	29.41	4.09	29.40	8.84	24.22	8.18	5.36
	2	29.45	4.36	28.64	3.10	1.82	1.77	1.77
696	1	23.33	2.98	23.32	7.02	24.21	7.49	4.20
	2	23.36	3.01	22.64	2.46	1.69	1.54	1.55
928	1	19.61	2.38	19.59	5.90	24.20	7.10	3.49
	2	19.61	2.35	18.96	2.07	1.61	1.39	1.40

Model does not fit (**Run 1**): Estimators $\hat{Y}_{dHT}$ and $\hat{Y}_{d\,CAL\,U}$ have the highest $\overline{CV}$ among all estimators; their $\overline{CV}$ s are quite comparable, implying that auxiliary data used at the population level in $\hat{Y}_{dCALU}$ has no impact on improving the reliability of the estimator at the domain level. The synthetic estimator $\hat{Y}_{dSYN}$ also has a high $\overline{CV}$ that remains constant no matter what the sample size is. The calibration estimator $\hat{Y}_{dCALU_d}$ has the lowest $\overline{CV}$ for all sample sizes. The ranking from low to high of the remaining three estimators is $\hat{Y}_{dPEBLUP}$, $\hat{Y}_{dEBLUP}$ and $\hat{Y}_{dREG}$. Note that the $\overline{CV}$ of $\hat{Y}_{dREG}$ decreases quite rapidly as compared to the other estimators. The reliability of all the estimators improves as the sample size increases.

Model fits (**Run 2**): The $\overline{CV}$s are smaller than those obtained in run 1 for all estimators except for $\hat{Y}_{dHT}$ and $\hat{Y}_{dCALU}$. This is expected as both estimators do not profit from the auxiliary data. These two estimators are still the ones with the highest $\overline{CV}$s. As expected, because the model fits well, all three small area estimators $\hat{Y}_{dSYN}$, $\hat{Y}_{dEBLUP}$ and $\hat{Y}_{dPEBLUP}$ have reasonable $\overline{CV}$s. The modified regression estimator $\hat{Y}_{dREG}$ performs better than the calibration at the domain level $\hat{Y}_{dCALU_d}$: the reverse was true when the model was incorrect (run 1).

Table 7. Average Relative Efficiency $\overline{RE}(\hat{Y}_{dEST})$

		Traditional Domain Estimators			**Small Area Estimators**			
Sample Size	**Run**	$\hat{Y}_{dHT}$	$\hat{Y}_{dCALU_d}$	$\hat{Y}_{dCALU}$	$\hat{Y}_{dREG}$	$\hat{Y}_{dSYN}$	$\hat{Y}_{dEBLUP}$	$\hat{Y}_{dPEBLUP}$
232	1	1.00	6.43	1.00	3.48	1.50	4.04	5.48
	2	1.00	6.57	1.03	8.48	13.23	13.97	13.96
464	1	1.00	7.39	1.00	3.47	1.03	3.25	5.59
	2	1.00	7.18	1.04	8.43	9.84	11.72	11.64
696	1	1.00	7.87	1.00	3.47	0.82	2.75	5.62
	2	1.00	7.85	1.04	8.41	8.03	10.67	10.56
928	1	1.00	8.07	1.00	3.46	0.69	2.40	5.64
	2	1.00	8.10	1.04	8.40	6.84	10.06	9.95

Note: The higher the number the more efficient the estimator relative to the HT estimator. Recall that run 1 represents the results when the model does not fit, whereas run 2 represents the results when the model fits.

Model does not fit (**Run 1**): The ranking of the estimators (from highest $\overline{RE}$ to lowest $\overline{RE}$) is as follows: $\hat{Y}_{d\,CALU_d}$, $\hat{Y}_{dPEBLUP}$, $\hat{Y}_{dEBLUP}$, $\hat{Y}_{d\,REG}$, $\hat{Y}_{dSYN}$, and $\hat{Y}_{d\,CALU}$. The traditional domain estimator $\hat{Y}_{d\,CALU_d}$ is doing the best, but it is closely followed by the two small area estimators $\hat{Y}_{dPEBLUP}$ and $\hat{Y}_{dEBLUP}$. As the sample size increases, there is a dichotomy in terms of $\overline{RE}$. The relative efficiency increases for $\hat{Y}_{d\,CALU_d}$ and $\hat{Y}_{dPEBLUP}$, whereas it decreases for $\hat{Y}_{d\,REG}$, $\hat{Y}_{dSYN}$, and $\hat{Y}_{dEBLUP}$. There is no change to $\hat{Y}_{d\,CALU}$ as the auxiliary information is not useful at the domain level.

Model fits (**Run 2**): The ranking of the estimators (from highest $\overline{RE}$ to lowest $\overline{RE}$) has changed with respect to run 1. It is now $\hat{Y}_{dEBLUP}$, $\hat{Y}_{dPEBLUP}$, $\hat{Y}_{dSYN}$, $\hat{Y}_{d\,REG}$, $\hat{Y}_{d\,CALU_d}$, and $\hat{Y}_{d\,CALU}$. The small area estimators are clearly more efficient than the traditional estimators. The relative efficiency increased for all estimators - maximum is now 14 versus 8 obtained in run 1. Once more, as the sample size increases, there is a dichotomy in terms of $\overline{RE}$.

Another way to summarize the behaviour of the various estimators is graphically. We summarized the average absolute relative bias, $ARB(\hat{Y}_{d\,EST})$, and the average coefficient of variation, $CV(\hat{Y}_{d\,EST})$, within each domain $d = 1, 2, \ldots, D,$ where $D = 29$.

Figures 2a and 2b display two typical graphs of the absolute relative bias over the domains for the two simulation runs. These graphs show the results for the sample size of 464. Similar results were obtained for the other sample sizes. We can see that the absolute relative bias of $\hat{Y}_{d\,SYN}$, $\hat{Y}_{d\,EBLUP}$ and $\hat{Y}_{d\,PEBLUP}$ is greatly reduced when we specify the 'correct' auxiliary variables in the underlying model. In the first run, the small area estimators show a 'drop' and a 'rise' between the groups of domains. This can be explained. The overall model fitted using $\boldsymbol{x}_{ij} = (1, x_{ij})$ produces a regression which is close to the underlying model for the second group of domains. Therefore, the differences are small for the second group of domains. However, this overall model is quite different from the one used to generate the population in the first and third groups of domains.

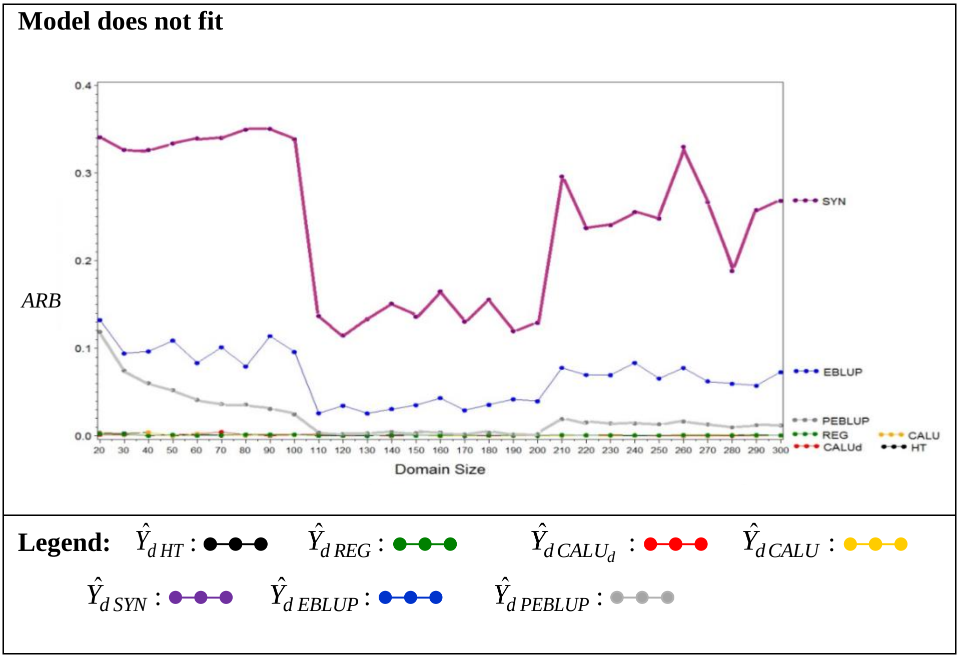

Figure 2a. Plots of the absolute relative bias of the estimators for sample size 464

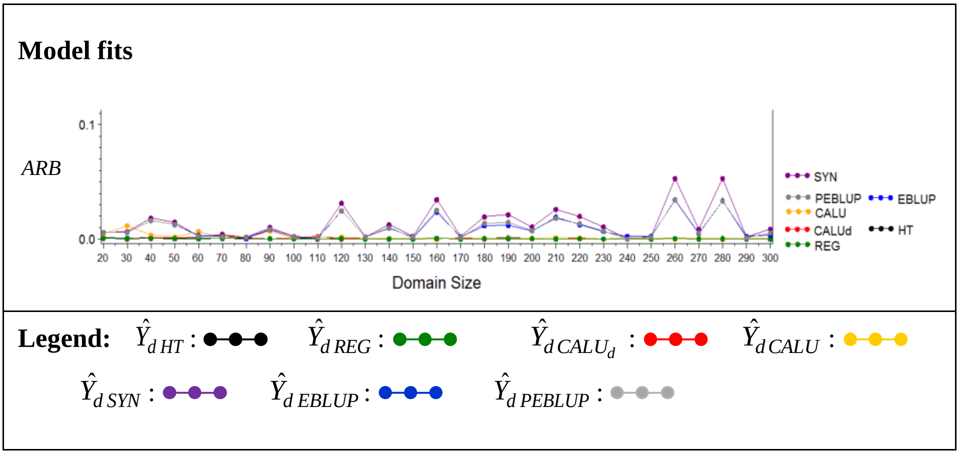

Figure 2b. Plots of the absolute relative bias of the estimators for sample size 464

Figures 3a and 3b display the coefficient of variation associated with the estimators. The coefficient of variation is reduced for all estimators except the HT estimator $\hat{Y}_{d\,HT}$ (which does not use any auxiliary information) and $\hat{Y}_{d\,CALU_d}$ (because the auxiliary variables for this estimator are equivalent in the two runs).

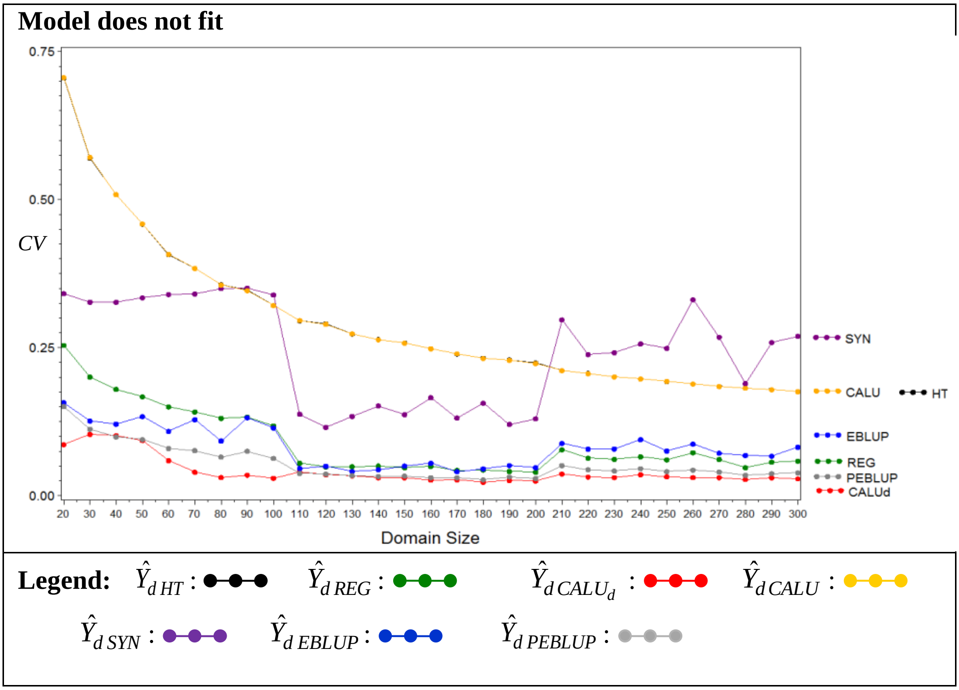

Figure 3a. Plots of the Coefficient of Variation of the Estimators for sample size 464

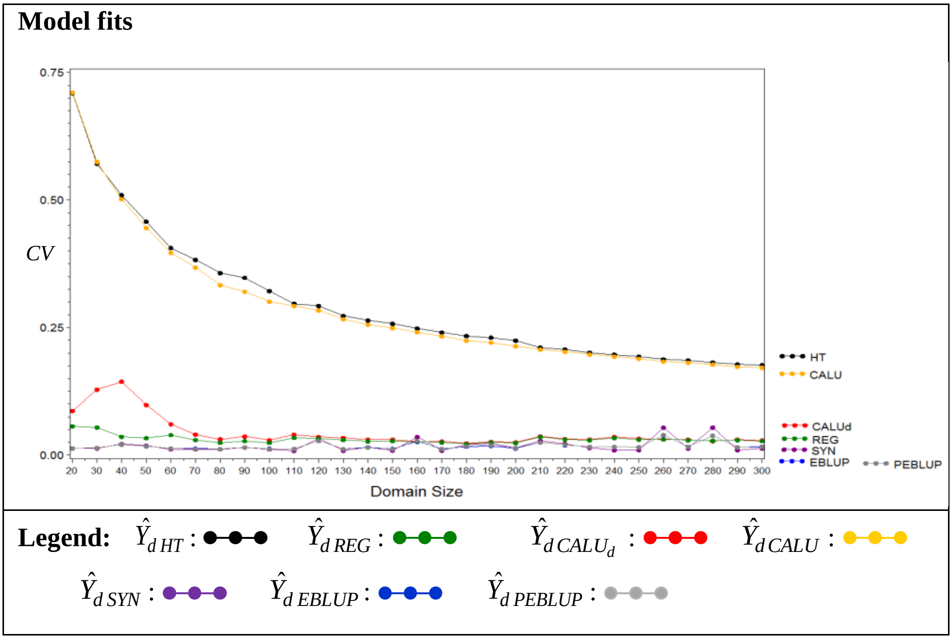

Figure 2b. Plots of the Coefficient of Variation of the Estimators for sample size 464

Figures 4a and 4b show a graphical display of the results for the average coefficient of variation $\overline{CV}(\hat{Y}_{EST})$ results given in Table 6. Under run 2, we see that $\hat{Y}_{d\,SYN}$, $\hat{Y}_{d\,EBLUP}$ and $\hat{Y}_{d\,PEBLUP}$ have the smallest $\overline{CV}(\hat{Y}_{EST})$. All three lines are indistinguishable as they are very close together. Under run 2, we see that $\hat{Y}_{d\,SYN}$, $\hat{Y}_{d\,EBLUP}$ and $\hat{Y}_{d\,PEBLUP}$ have the smallest $\overline{CV}(\hat{Y}_{EST})$. All three lines are indistinguishable as they are very close together.

Model does not fit

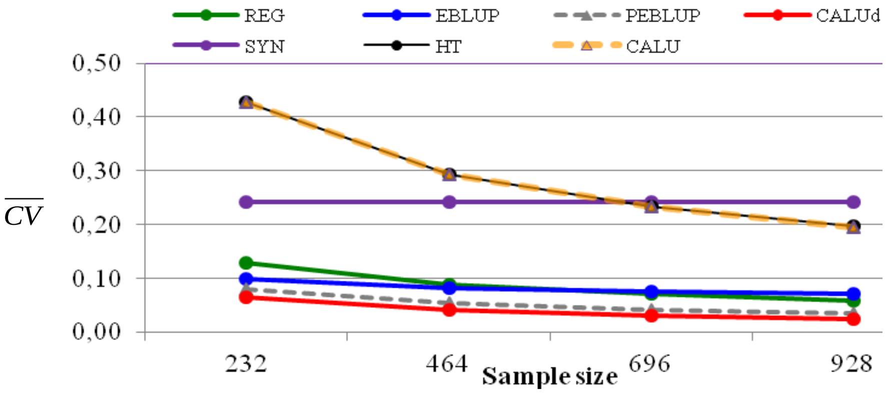

Figure 4a. Plots of the average coefficient of variation of the estimators by sample size

Model fits

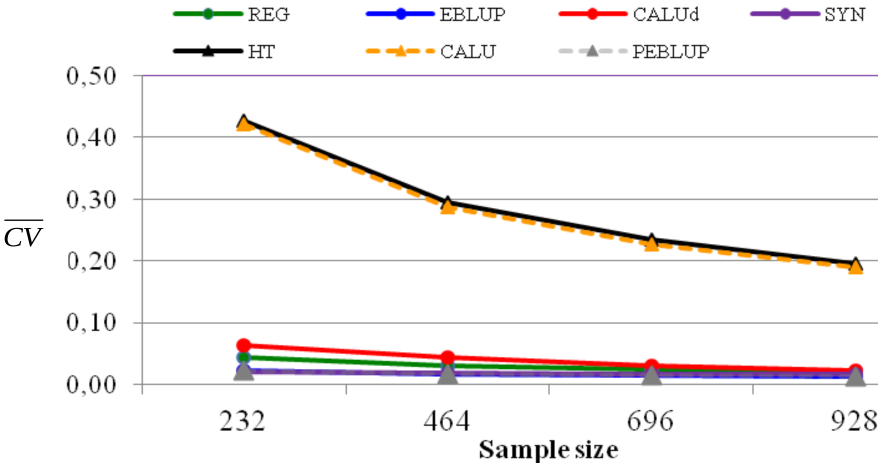

Figure 4b. Plots of the average coefficient of variation of the estimators by sample size

Figures 5a and 5b show a graphical display of the average relative efficiency of the estimators given in table 7. Under run 2, we note that $\hat{Y}_{d\ EBLUP}$ and $\hat{Y}_{d\ PEBLUP}$ have the highest $\overline{RE}(\hat{Y}_{EST})$ over the various sample sizes.

Model does not fit

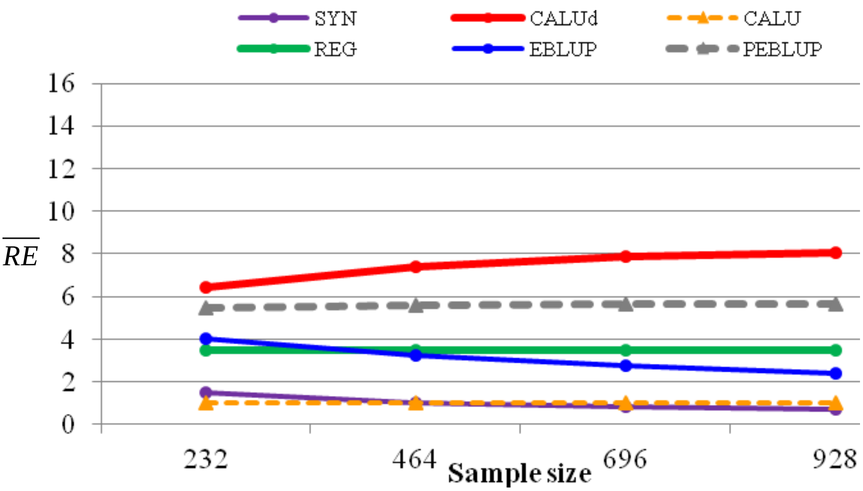

Figure 5a. Plots of the average relative efficiency of the estimators by sample size

Model fits

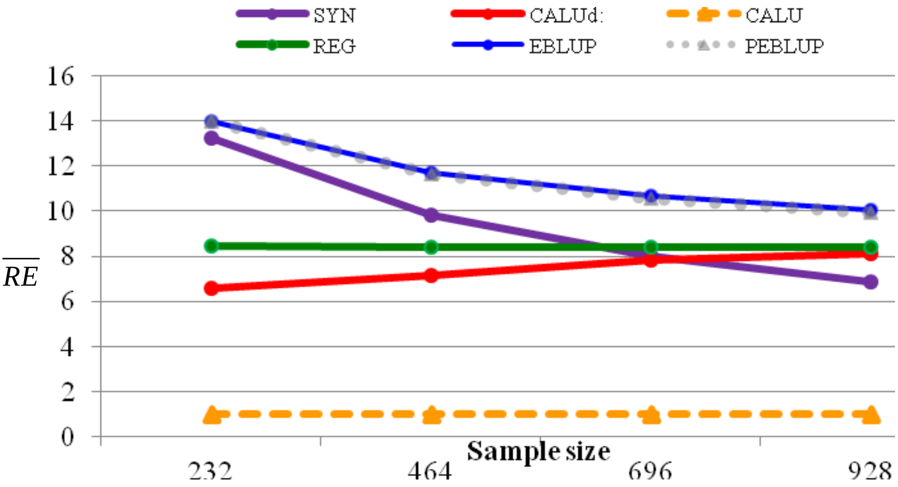

Figure 5b. Plots of the average relative efficiency of the estimators by sample size

4. Conclusions

We compared via simulation the behavior of a number of traditional domain and small area estimators. The sampling design used in the simulation, simple random sampling without replacement, is a simplification of the sampling design commonly used for business surveys (stratified simple random sampling without replacement). The estimators using the auxiliary data either reflected the model used to generate the population (model fits) or did not (model does not fit). The simulation design did not use unequal probability sampling. The additional complexity of using unequal probability sampling is that we would have had to modify our model-based small area estimators to account for possible informative sampling. However, since we used simple random sampling without replacement, we did not have to account for this problem.

The conclusions of our simulation are as follows. Comparing the efficiency between the traditional and small area estimators, the results very much depend on whether the model holds or not. The calibration estimator $\hat{Y}_{d\,CAL\,U}$ which only uses auxiliary data at the population level is not efficient at the domain level whether the model holds or not. This is in contrast to $\hat{Y}_{d\,CAL\,U_d}$ that uses auxiliary data at the domain level. The estimator $\hat{Y}_{d\,CAL\,U_d}$ is the best traditional estimator to use when the model holds. Its average relative efficiency increases as the overall sample size increases. Its weakness is in the smaller domains, where the expected sample size is smaller than three units, as it cannot be defined when the auxiliary data consists of two auxiliary variables; in general, when there are p auxiliary variables, we are not be able to define $\hat{Y}_{d\,CAL\,U_d}$ when the sample size is smaller than p+1 auxiliary variables. When the model does not hold, $\hat{Y}_{d\,REG}$ is the best traditional estimator to use. However, it is outperformed by the small area estimators $\hat{Y}_{d\,SYN}$, $\hat{Y}_{d\,EBLUP}$ and $\hat{Y}_{d\,PEBLUP}$. The small area estimator $\hat{Y}_{d\,EBLUP}$ is the most efficient one when the model holds, although it is closely followed by $\hat{Y}_{d\,SYN}$ and $\hat{Y}_{d\,PEBLUP}$. When the model does not hold, the $\hat{Y}_{d\,PEBLUP}$ estimator is the most efficient small area estimator; an explanation for this is that it is design-consistent.

Acknowledgement

We would like to thank the referee for his constructive comments that substantially improved this paper.

REFERENCES

BATTESE, G. E., HARTER, R. M., 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 (401), 28–36.

DATTA, G. S., (2009). Model-based approach to small area estimation. *Handbook of Statistics*, 29, 251–288.

DEVILLE, J. C., SÄRNDAL, C. E., (1992). Calibration estimation in survey sampling. *Journal* of *the American Statistical Association*, 87(418), 376–382.

ESTEVAO, V., HIDIROGLOU, M. A., YOU, Y., (2014). Methodology Software Library - Small area Estimation Methodology Specifications for Area and Unit Level based Models. Technical Report, Statistics Canada.

FAY, R. E., HERRIOT, R. A., (1979). Estimation of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74 (366A), 269–277.

LEHTONEN, R., VEIJANEN, A., (2009). Design-based methods of estimation for domains and small areas. *Handbook of statistics*, 29, 219–249.

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

PFEFFERMANN, D., SVERCHKOV, M., (2007). Small Area Estimation Under Informative Probability Sampling of Areas and Within the Selected Areas. *Journal of the American Statistical Association* 102 (480), 1427–1439.

RAO, J. N. K., (2003). *Small Area Estimation:* John Wiley & Sons.

SINGH, A. C., MIAN, I. U. H., (1995). Generalized Sample Size Dependent Estimators for Small Areas, *Proceedings of the 1995 Annual Research Conference,* U.S. Bureau of the Census, Washington, DC, 687–701.

SINGH, M. P., GAMBINO, J., MANTEL, H., (1994). Issues and Strategies for Small Area Data. *Survey Methodology*, 20 (1), 3–22.

WOODRUFF, R. S., (1966). Use of a Regression Technique to Produce Area Breakdowns of the Monthly National Estimates of Retail Trade. *Journal* of *the American Statistical Association*, 61 (314), 496–504.

YOU, Y., RAO, J. N. K., (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights, *Canadian Journal* of *Statistics,* 30, 431–439**.**

VERRET, F., RAO, J. N. K., VERRET, F., RAO, J. N. K., HIDIROGLOU, M. A., (2015). Model-based small area estimation under informative sampling. To appear in the December 2015 issue of *Survey Methodology*.