A Comparison of Small Area and Traditional Estimators via Simulation

M.A. Hidiroglou and V.M. Estevao
Statistical Research and Innovation Division, Statistics Canada,
16th Floor, R. H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6

1 Introduction

Domain estimates at Statistics Canada are typically obtained using well-established methods based on calibration estimation. 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 $\boldsymbol{x}$ available from administrative files. Small area estimation essentially combines the associated 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.

A SAS based prototype (Estevao et al 2014a and 2014b) has been recently developed at Statistics Canada to produce small area estimates. It 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.

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 include 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 setup for the simulations is described in section 3. Section 4 presents the measures used to assess the results. The study and its results are given in section 5. Section 6 provides a few conclusions from our findings.

2 Estimators Studied in the Simulations

The main objective of the study is to compare the properties of model-based small-area estimators with those traditionally used in survey estimation. In our simulations, we draw repeated samples s from U using simple random sampling without replacement. The portion of the sample in domain U_d is denoted by s_d. The number of units in s_d is given by n_d.

We considered seven estimators of a domain total $Y_d = \sum_{j \in U_d} y_{dj}$. To simplify the presentation, we divided the estimators into the two groups shown in Table 1 and Table 2. Table 1 shows the traditional estimators and Table 2 shows the small-area estimators.

We use the following notation in both tables. The survey design weight for unit j in domain i is denoted by w_{ij}. The corresponding value of the variable of interest is given by y_{ij}. The vector of auxiliary variables is shown as $\boldsymbol{x}_{ij}$. Other terms are explained in later sections.

Table 1: Traditional Domain Estimators

Estimator	Formula	Properties
Horvitz-Thompson (HT)	$\hat{Y}_{d\,HT} = \begin{cases} \sum_{j\in s_d} w_{dj}\, y_{dj} & \text{if } n_d > 0 \\ 0 & \text{if } n_d = 0 \end{cases}$	*design-unbiased*
Calibration at the domain level (CALUd)	$\hat{Y}_{d\,CALU_d} = \begin{cases} \sum_{j\in s_d} w_{dj}\, y_{dj} + (\boldsymbol{X}_d - \hat{\boldsymbol{X}}_{d\,HT})^T\, \hat{\boldsymbol{\beta}}_{d\,CALU_d} & \text{if } n_d \geq 3 \\ .\ \text{(missing)} & \text{if } n_d < 3 \end{cases}$ with $\boldsymbol{X}_d = \sum_{j\in U_d} \boldsymbol{x}_{dj}$, $\hat{\boldsymbol{X}}_{d\,HT} = \sum_{j\in s_d} w_{dj}\, \boldsymbol{x}_{dj}$ and $\hat{\boldsymbol{\beta}}_{d\,CALU_d} = \left(\sum_{j\in s_d} \frac{w_{dj}\, \boldsymbol{x}_{dj}\, \boldsymbol{x}_{dj}^T}{c_{dj}} \right)^{-1} \sum_{j\in s_d} \frac{w_{dj}\, \boldsymbol{x}_{dj}\, y_{dj}}{c_{dj}}$	*approximately design-unbiased* if the expected domain size is large
Calibration at the population level (CALU)	$\hat{Y}_{d\,CALU} = \begin{cases} \sum_{j\in s_d} w_{dj}\, y_{dj} + (\boldsymbol{X} - \hat{\boldsymbol{X}}_{HT})^T\, \hat{\boldsymbol{\beta}}_{CALU} & \text{if } n_d > 0 \\ 0 & \text{if } n_d = 0 \end{cases}$ with $\boldsymbol{X} = \sum_{i=1}^{D} \sum_{j\in U_i} \boldsymbol{x}_{ij}$, $\hat{\boldsymbol{X}}_{HT} = \sum_{i=1}^{D} \sum_{j\in s_i} w_{ij}\, \boldsymbol{x}_{ij}$ and $\hat{\boldsymbol{\beta}}_{CALU} = \left(\sum_{i=1}^{D} \sum_{j\in s_i} \frac{w_{ij}\, \boldsymbol{x}_{ij}\, \boldsymbol{x}_{ij}^T}{c_{ij}} \right)^{-1} \sum_{j\in s_d} \frac{w_{dj}\, \boldsymbol{x}_{dj}\, y_{dj}}{c_{dj}}$	*approximately design-unbiased* if the expected domain size is large
Modified regression (REG)	$\hat{Y}_{d\,REG} = \begin{cases} \sum_{j\in s_d} w_{dj}\, y_{dj} + (\boldsymbol{X}_d - \hat{\boldsymbol{X}}_{d\,HT})^T\, \hat{\boldsymbol{\beta}}_{REG} & \text{if } n_d > 0 \\ \boldsymbol{X}_d^T\, \hat{\boldsymbol{\beta}}_{REG} & \text{if } n_d = 0 \end{cases}$ with $\boldsymbol{X}_d = \sum_{j\in U_d} \boldsymbol{x}_{dj}$, $\hat{\boldsymbol{X}}_{d\,HT} = \sum_{j\in s_d} w_{dj}\, \boldsymbol{x}_{dj}$ and $\hat{\boldsymbol{\beta}}_{REG} = \left(\sum_{i=1}^{D} \sum_{j\in s_i} \frac{w_{ij}\, \boldsymbol{x}_{ij}\, \boldsymbol{x}_{ij}^T}{c_{ij}} \right)^{-1} \sum_{i=1}^{D} \sum_{j\in s_i} \frac{w_{ij}\, \boldsymbol{x}_{ij}\, y_{ij}}{c_{ij}}$	*design-unbiased* as the overall sample size increases
Synthetic (SYN)	$\hat{Y}_{d\,SYN} = \boldsymbol{X}_d^T\, \hat{\boldsymbol{\beta}}_{SYN}$ with $\boldsymbol{X}_d = \sum_{j\in U_d} \boldsymbol{x}_{dj}$ and $\hat{\boldsymbol{\beta}}_{SYN} = \left(\sum_{i=1}^{D} \sum_{j\in s_i} \frac{w_{ij}\, \boldsymbol{x}_{ij}\, \boldsymbol{x}_{ij}^T}{c_{ij}} \right)^{-1} \sum_{i=1}^{D} \sum_{j\in s_i} \frac{w_{ij}\, \boldsymbol{x}_{ij}\, y_{ij}}{c_{ij}}$	*design-biased* with $Bias(\hat{Y}_{d\,SYN}) \doteq \boldsymbol{X}_d^T\, \boldsymbol{B} - Y_d$ where $\boldsymbol{B}$ is the population regression vector

Table 2: Small-Area Estimators

Estimator	Formula
EBLUP (EBLUP)	$\hat{Y}_{d\,EBLUP} = \begin{cases} N_d\,\{\bar{\boldsymbol{X}}_d^T\,\hat{\boldsymbol{\beta}}_{EBLUP} + \hat{\gamma}_{da}(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T\,\hat{\boldsymbol{\beta}}_{EBLUP})\} & \text{if } n_d > 0 \\ \boldsymbol{X}_d^T\,\hat{\boldsymbol{\beta}}_{EBLUP} & \text{if } n_d = 0 \end{cases}$ with $\bar{y}_{da} = \dfrac{\sum_{j\in s_d} a_{dj}y_{dj}}{\sum_{j\in s_d} a_{dj}}$ and $\bar{\boldsymbol{x}}_{da} = \dfrac{\sum_{j\in s_d} a_{dj}\boldsymbol{x}_{dj}}{\sum_{j\in s_d} a_{dj}}$ $\hat{\gamma}_{da} = \dfrac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2\,\delta_{da}^2}$ with $\delta_{da}^2 = \dfrac{1}{\sum_{j\in s_d} a_{dj}}$ for $d = 1,2,\ldots,D$ and $\hat{\boldsymbol{\beta}}_{EBLUP} = \left(\sum_{i=1}^{D}\sum_{j\in s_i} a_{ij}(\boldsymbol{x}_{ij} - \hat{\gamma}_{ia}\,\bar{\boldsymbol{x}}_{ia})\,\boldsymbol{x}_{ij}^T\right)^{-1}\sum_{i=1}^{D}\sum_{j\in s_i} a_{ij}(\boldsymbol{x}_{ij} - \hat{\gamma}_{ia}\,\bar{\boldsymbol{x}}_{ia})\,y_{ij}$
Pseudo-EBLUP (PEBLUP)	$\hat{Y}_{d\,PEBLUP} = \begin{cases} N_d\,\{\bar{\boldsymbol{X}}_d^T\,\hat{\boldsymbol{\beta}}_{PEBLUP} + \hat{\gamma}_{dw}(\bar{y}_{dw} - \bar{\boldsymbol{x}}_{dw}^T\,\hat{\boldsymbol{\beta}}_{PEBLUP})\} & \text{if } n_d > 0 \\ \boldsymbol{X}_d^T\,\hat{\boldsymbol{\beta}}_{PEBLUP} & \text{if } n_d = 0 \end{cases}$ with $\bar{y}_{dw} = \dfrac{\sum_{j\in s_d} w_{dj}y_{dj}}{\sum_{j\in s_d} w_{dj}}$, $\bar{\boldsymbol{x}}_{dw} = \dfrac{\sum_{j\in s_d} w_{dj}\boldsymbol{x}_{dj}}{\sum_{j\in s_d} w_{dj}}$ $\hat{\gamma}_{dw} = \dfrac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2\,\delta_{dw}^2}$ where $\delta_{dw}^2 = \dfrac{\sum_{j\in s_d} w_{dj}^2 \big/ a_{dj}}{\left(\sum_{j\in s_d} w_{dj}\right)^2}$ for $d = 1,2,\ldots,D$, $\hat{\boldsymbol{\beta}}_{PEBLUP} = \left(\sum_{i=1}^{D}\sum_{j\in s_i} w_{ij}\,a_{ij}(\boldsymbol{x}_{ij} - \hat{\gamma}_{iwa}\,\bar{\boldsymbol{x}}_{iwa})\,\boldsymbol{x}_{ij}^T\right)^{-1}\sum_{i=1}^{D}\sum_{j\in s_i} w_{ij}\,a_{ij}(\boldsymbol{x}_{ij} - \hat{\gamma}_{iwa}\,\bar{\boldsymbol{x}}_{iwa})\,y_{ij}$ with $\bar{\boldsymbol{x}}_{iwa} = \dfrac{\sum_{j\in s_i} w_{ij}\,a_{ij}\,\boldsymbol{x}_{ij}}{\sum_{j\in s_i} w_{ij}\,a_{ij}}$ and $\hat{\gamma}_{iwa} = \dfrac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2\,\delta_{iwa}^2}$ where $\delta_{iwa}^2 = \dfrac{\sum_{j\in s_i} (w_{ij}\,a_{ij})^2 \big/ a_{ij}}{\left(\sum_{j\in s_i} w_{ij}\,a_{ij}\right)^2}$ for $i = 1,2,\ldots,D$

Note: $a_{ij} = 1\big/k_{ij}^2$

The estimators in Table 1 are the Horvitz-Thompson estimator $\hat{Y}_{d\,HT}$, two calibration estimators, $\hat{Y}_{d\,CALU_d}$ and $\hat{Y}_{d\,CALU}$ (inspired by Deville and Särndal 1982), based on auxiliary information at the domain and population respectively, the modified regression estimator $\hat{Y}_{d\,REG}$ and the synthetic estimator $\hat{Y}_{d\,SYN}$.

The estimator $\hat{Y}_{d\,HT}$ uses no auxiliary information. This estimator is unbiased but produces inefficient estimates compared to the others. When there are no sample units in the domain, we set $\hat{Y}_{d\,HT}$ to 0. This ensures that the estimator is unbiased for Y_d over all samples.

The two calibration estimators use auxiliary information at different levels. Estimator $\hat{Y}_{d\,CALU_d}$ uses auxiliary information at the domain level while $\hat{Y}_{d\,CALU}$ uses information at the population level. The estimator $\hat{Y}_{d\,CALU_d}$ is known to be more efficient than $\hat{Y}_{d\,CALU}$. However, $\hat{Y}_{d\,CALU_d}$ 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 $\hat{Y}_{d\,CALU_d}$ 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. In our simulations we use two auxiliary variables; one of these has a constant value of 1 and represents the intercept in the model. As a minimal requirement, we produce $\hat{Y}_{d\,CALU_d}$ only if there are 3 or more units in a domain. Otherwise, we cannot produce a value, so we set it to missing. This means that we only work with a subset of all possible samples. Earlier simulations showed that we cannot set the value to 0 as this would result in a biased estimator. As for $\hat{Y}_{d\,CALU}$, when there are no sample units in the domain, we set the value of this estimator to 0. This ensures that the estimator is approximately design unbiased for the domain total.

The synthetic estimator (Gonzalez 1973) was used to produce estimates for small areas before the development of the EBLUP and Pseudo-EBLUP estimators. The synthetic estimator uses the same regression coefficient as the modified regression estimator ($\hat{\boldsymbol{\beta}}_{SYN} = \hat{\boldsymbol{\beta}}_{REG}$).

The modified regression estimator $\hat{Y}_{d\,REG}$ (Woodruff 1966) is a hybrid of $\hat{Y}_{d\,CALU_d}$ and $\hat{Y}_{d\,CALU}$. It requires auxiliary totals at the domain level but makes use a regression coefficient at the population level. When there are no sample units in the domain, we produce the synthetic estimate given by $\boldsymbol{X}_d^T\,\hat{\boldsymbol{\beta}}_{REG} = \boldsymbol{X}_d^T\,\hat{\boldsymbol{\beta}}_{SYN}$.

The small-area estimators in Table 2 are based on a hierarchical model at the unit level. This model is as follows.

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i + e_{ij}\text{, where } v_i \sim iid(0,\sigma_v^2) \text{ and } e_{ij} \sim iid(0,k_{ij}^2\,\sigma_e^2) \qquad (1)$$

In our application of this model, the areas are our domains of interest. The quantity $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ is the fixed effect which is assumed to be a linear combination of the auxiliary variables $\boldsymbol{x}_{ij}$. The term v_i is a random effect for the area (domain) and the e_{ij} are the random errors at the unit level. The term k_{ij}^2 is a coefficient which reflects the heterogeneity of the random errors. This coefficient translates to a_{ij} in the various formulas in Table 2 through the definition $a_{ij} = 1\big/k_{ij}^2$.

The small area estimator $\hat{Y}_{d\,EBLUP}$, given in Rao (2003, p.136), is an extension of the Battese, Harter, and Fuller (1988) for the case where the error structure of the residuals is not homogeneous. Estimator $\hat{Y}_{d\,PEBLUP}$ is an extension of the Pseudo-EBLUP estimator given in You and Rao (2002) that accounts for the heterogeneity of the e_{ij} residuals in the model given by (1). It includes the survey weights w_{ij} in the regression coefficient and the parameter estimate.

3 The Set Up for the Populations and Domains in the Simulations

For each of our simulations, we created a population U of size N with $D = 29$ mutually exclusive domains U_d $(d = 1,\ldots,D)$. Each domain had a different number of units N_d. These numbers were equally spaced with $N_1 = 20$, $N_2 = 30$, $N_3 = 40$, all the way up to domain U_{29} with $N_{29} = 300$ units. This means $N = \sum_{d=1}^{D} N_d = 4640$.

Each simulation involved the selection of 100,000 independent samples and the calculation 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 many surveys.

The sample units in domain U_d are denoted by s_d with $s = \bigcup_{d=1}^{D} s_d$. We observe 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 $n_d = 0$, $n_d = 1$ and $n_d = 2$ in the three smallest domains when the sample size $n = 232$.

Table 3: Hypergeometric Probabilities in the Three Smallest Domains when Sample Size $n = 232$

	U_1 with $N_1 = 20$	U_2 with $N_2 = 30$	U_3 with $N_3 = 40$
$Prob(n_d = 0)$	0.3577119	0.2135777	0.1273735
$Prob(n_d = 1)$	0.3781689	0.3394612	0.2705485
$Prob(n_d = 2)$	0.1890414	0.2595948	0.2788755

From this table, we can see what happens in the smallest domain (U_1) when the sample size $n = 232$ (with the probabilities highlighted in yellow). We produce $\hat{Y}_{d\,HT} = 0$ and $\hat{Y}_{d\,CALU} = 0$ about 36% of the time (whenever $n_d = 0$). Since we require $n_d \ge 3$, we cannot produce an estimate for $\hat{Y}_{d\,CAL\,U_d}$ in approximately 92.5% of the samples.

4 Measures Used in the Analysis of the Simulation Results

In each simulation, we selected $R = 100{,}000$ independent SRWOR samples. For each sample, we calculated estimates of Y_d based on the seven estimators. This allowed us to produce the simulation bias, variance and mean squared error. Their definitions are given below.

$$Bias(\hat{Y}_{d\,EST}) = \frac{\sum_{r=1}^{R} \hat{Y}_{d\,EST}^{(r)}}{R} - Y_d,\ Var(\hat{Y}_{d\,EST}) = \frac{\sum_{r=1}^{R} \left(\hat{Y}_{d\,EST}^{(r)} - \left[\sum_{r=1}^{R} \hat{Y}_{d\,EST}^{(r)} \Big/ R \right] \right)^2}{R},\ MSE(\hat{Y}_{d\,EST}) = \frac{\sum_{r=1}^{R} (\hat{Y}_{d\,EST}^{(r)} - Y_d)^2}{R}$$

We use $\hat{Y}_{d\,EST}^{(r)}$ to denote 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.

In addition to these estimates for each domain, we included 3 measures to summarize the properties of each estimator over all of the domains.

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

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

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

The measure given by $\overline{RE}(\hat{Y}_{EST})$ 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.

5 Simulations

We looked at the effect of fitting an improper model. We did this by generating a population under a specific model and then specifying a set of auxiliary variables that did not reflect the model.

We created values for the single auxiliary *x* using a $Gamma(\alpha, \beta)$ distribution. Then we generated values for the variable of interest *y* through the model specified below for the units in the domains U_i for $i = 1, 2, ..., 29$.

$$y_{ij} = b_0 + b_1 x_{ij} + v_i + e_{ij}, \text{ where } v_i = Normal(0, \sigma_v^2) \text{ and } e_{ij} = Normal(0, k_{ij}^2 \sigma_e^2)$$
$$\text{with } x_{ij} = Gamma(5,10)$$

The auxiliary variable *x* has a $Gamma(\alpha = 5, \beta = 10)$ distribution with mean $\alpha\beta = 50$ and variance $\alpha\beta^2 = 500$. We used $\sigma_v^2 = \sigma_e^2 = 20^2 = 400$ and set $k_{ij}^2 = x_{ij}$.

We placed the domains into 3 groups and used a different intercept b_0 and slope b_1 for each group. Our choices were as follows.

Group	Domains in Group	b_0	b_1
1	U_i for $i = 1...9$	200	30
2	U_i for $i = 10...19$	300	20
3	U_i for $i = 20...29$	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 regression lines (with the corresponding b_0 and b_1) are shown in red. Without the colours to identify the groups (as shown in the inset of Figure 1), one might be inclined to think that the population was generated under a simple model with a single auxiliary variable (one intercept and slope). Therefore, in the first run (run 1) of our study, we produced the seven estimators using the auxiliary variables $\boldsymbol{x}_{ij} = (1, x_{ij})$. In the second run (run 2), we acknowledged that there are three separate models and used a set of auxiliary variables reflecting the manner in which the population values were generated. This meant using a set of dummy-coded auxiliary variables defined as follows for each unit:

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

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

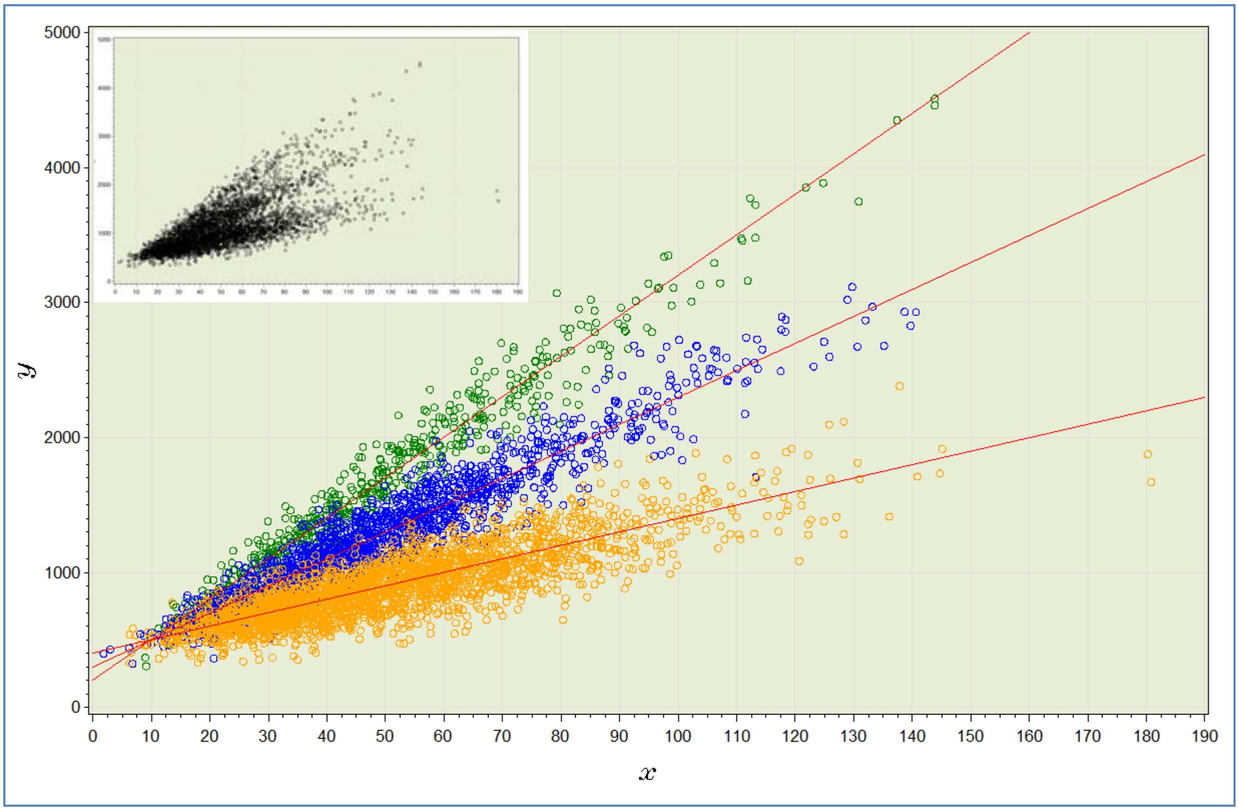

Figure 1: Plot of *y* vs *x* for Population in the Simulation Study

We carried out two separate runs with the two different sets of auxiliary variables. In each run, we selected 100,000 independent samples from the population and produced estimates of the domain totals Y_d based on the seven estimators. Table 4, Table 5 and Table 6 show the differences between the two runs using the summary statistics described in the previous section. The results are presented after these three tables.

Table 4: Average Absolute Relative Bias $\overline{ARB}(\hat{Y}_{EST})$

Sample Size	Run	$\hat{Y}_{HT}$	$\hat{Y}_{SYN}$	$\hat{Y}_{CALU_d}$	$\hat{Y}_{CALU}$	$\hat{Y}_{REG}$	$\hat{Y}_{EBLUP}$	$\hat{Y}_{PEBLUP}$
232	**1**	0.115	24.178	0.161	0.149	0.188	7.582	4.115
	2	0.110	1.332	0.152	0.356	0.048	1.066	1.081
464	**1**	0.083	24.179	0.082	0.089	0.096	6.709	2.240
	2	0.060	1.332	0.073	0.186	0.023	0.945	0.963
696	**1**	0.062	24.179	0.046	0.061	0.062	6.426	1.519
	2	0.045	1.333	0.044	0.110	0.017	0.844	0.861
928	**1**	0.057	24.179	0.032	0.059	0.044	6.289	1.137
	2	0.052	1.333	0.032	0.086	0.011	0.761	0.772

Note: All numbers are percentages. Lower numbers indicate more accurate estimators.

Table 5: Average Coefficient of Variation $\overline{CV}(\hat{Y}_{EST})$

Sample Size	Run	$\hat{Y}_{HT}$	$\hat{Y}_{SYN}$	$\hat{Y}_{CALU_d}$	$\hat{Y}_{CALU}$	$\hat{Y}_{REG}$	$\hat{Y}_{EBLUP}$	$\hat{Y}_{PEBLUP}$
232	**1**	42.787	24.268	6.565	42.809	12.817	9.895	7.926
	2	42.765	2.166	6.389	42.044	4.467	2.218	2.213
464	**1**	29.408	24.221	4.094	29.403	8.839	8.183	5.359
	2	29.446	1.823	4.357	28.638	3.097	1.766	1.769
696	**1**	23.333	24.205	2.977	23.319	7.018	7.487	4.196
	2	23.358	1.686	3.006	22.636	2.463	1.541	1.548
928	**1**	19.609	24.198	2.380	19.594	5.901	7.100	3.494
	2	19.606	1.607	2.351	18.963	2.071	1.386	1.395

Note: All numbers are percentages. Lower numbers indicate more efficient estimators.

Table 6: Average Relative Efficiency $\overline{RE}(\hat{Y}_{EST})$

Sample Size	Run	$\hat{Y}_{HT}$	$\hat{Y}_{SYN}$	$\hat{Y}_{CALU_d}$	$\hat{Y}_{CALU}$	$\hat{Y}_{REG}$	$\hat{Y}_{EBLUP}$	$\hat{Y}_{PEBLUP}$
232	**1**	1.000	1.498	6.434	1.000	3.476	4.044	5.479
	2	1.000	13.227	6.571	1.028	8.477	13.965	13.958
464	**1**	1.000	1.032	7.393	1.001	3.465	3.245	5.590
	2	1.000	9.840	7.176	1.035	8.431	11.720	11.644
696	**1**	1.000	0.819	7.869	1.001	3.465	2.747	5.621
	2	1.000	8.027	7.846	1.037	8.413	10.665	10.561
928	**1**	1.000	0.689	8.068	1.001	3.462	2.395	5.639
	2	1.000	6.842	8.098	1.039	8.403	10.059	9.950

Note: The higher the number the more efficient the estimator relative to the HT estimator.

Run 1 Results

- The calibration estimator $\hat{Y}_{dCALU_d}$ is generally the best estimator for most domains and sample sizes. This is certainly true for large domains and large sample sizes. However, this estimator is very sensitive to the realized sample size in the domain (n_d). For small domains, this estimator can produce erratic and wide ranging estimates when the realized sample size in the domain is small.

- On balance, the results show that when we specify a set of auxiliary variables that do not adequately reflect the underlying model for the population, the estimator $\hat{Y}_{d\,CALU_d}$ provides the best protection against model misspecification. This makes sense because the resulting calibration at the domain level leads to a separate fit for each domain which is known to be more efficient.
- The two small-area estimators $\hat{Y}_{d\,EBLUP}$ and $\hat{Y}_{d\,PEBLUP}$ produce efficient estimates for small domains and small sample sizes but they are not as efficient as $\hat{Y}_{d\,CALU_d}$ for larger domains and sample sizes.
- The synthetic estimator $\hat{Y}_{d\,SYN}$ and the two small-area estimators have larger bias than the other estimators. This is much more pronounced for the synthetic estimator than the small-area estimators.
- The synthetic estimator is not sensitive to the realized sample size in the domain. We do not see a significant reduction in the MSE as we increase the overall sample size. This is true for all domains.
- The estimators $\hat{Y}_{d\,HT}$ and $\hat{Y}_{d\,CALU}$ have similar properties across all domains and sample sizes. They tend to produce the most inefficient estimates in terms of MSE. The estimator $\hat{Y}_{d\,CALU}$ uses auxiliary information at the population level to produce estimates at the domain level. This is very inefficient when the domains are considerably smaller than the population. In fact, the results suggest that this is almost equivalent to using no auxiliary information.
- The modified regression estimator $\hat{Y}_{d\,REG}$ performs better than $\hat{Y}_{d\,HT}$ and $\hat{Y}_{d\,CALU}$ but is less efficient than $\hat{Y}_{d\,CALU_d}$.

Run 2 Results

- With the dummy-coded auxiliary variables we see an improvement in all estimators except $\hat{Y}_{d\,HT}$.
- The greatest improvement is seen for both $\hat{Y}_{d\,SYN}$ and the small-areas estimators $\hat{Y}_{d\,EBLUP}$ and $\hat{Y}_{d\,PEBLUP}$. This is expected because we now have the 'right' model. These estimators produce the most efficient estimates for all domains and sample sizes.
- For $\hat{Y}_{d\,CALU}$ we see only a small improvement in efficiency from the first run.
- The modified regression estimator $\hat{Y}_{d\,REG}$ is not as efficient as the synthetic and small-area estimators but it is better than the other estimators including $\hat{Y}_{d\,CALU_d}$.
- There is no real change in $\hat{Y}_{d\,CALU_d}$ from the first run. This is evident when the domain size and sample size are sufficient large. Otherwise, this is not apparent because of the wide variability of the estimator when the realized sample size in the domain is small. The reason why we should not see a change in the estimator is as follows. The dummy coded auxiliary variables in the second run produce the same calibration constraints at the domain level as the auxiliary variables in the first run. Therefore, there is no change in the definition of the estimator. The differences that we see between the first and second runs are simply due to the sampling variability between the two different sets of 100,000 independent samples.

Figure 2 shows 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.

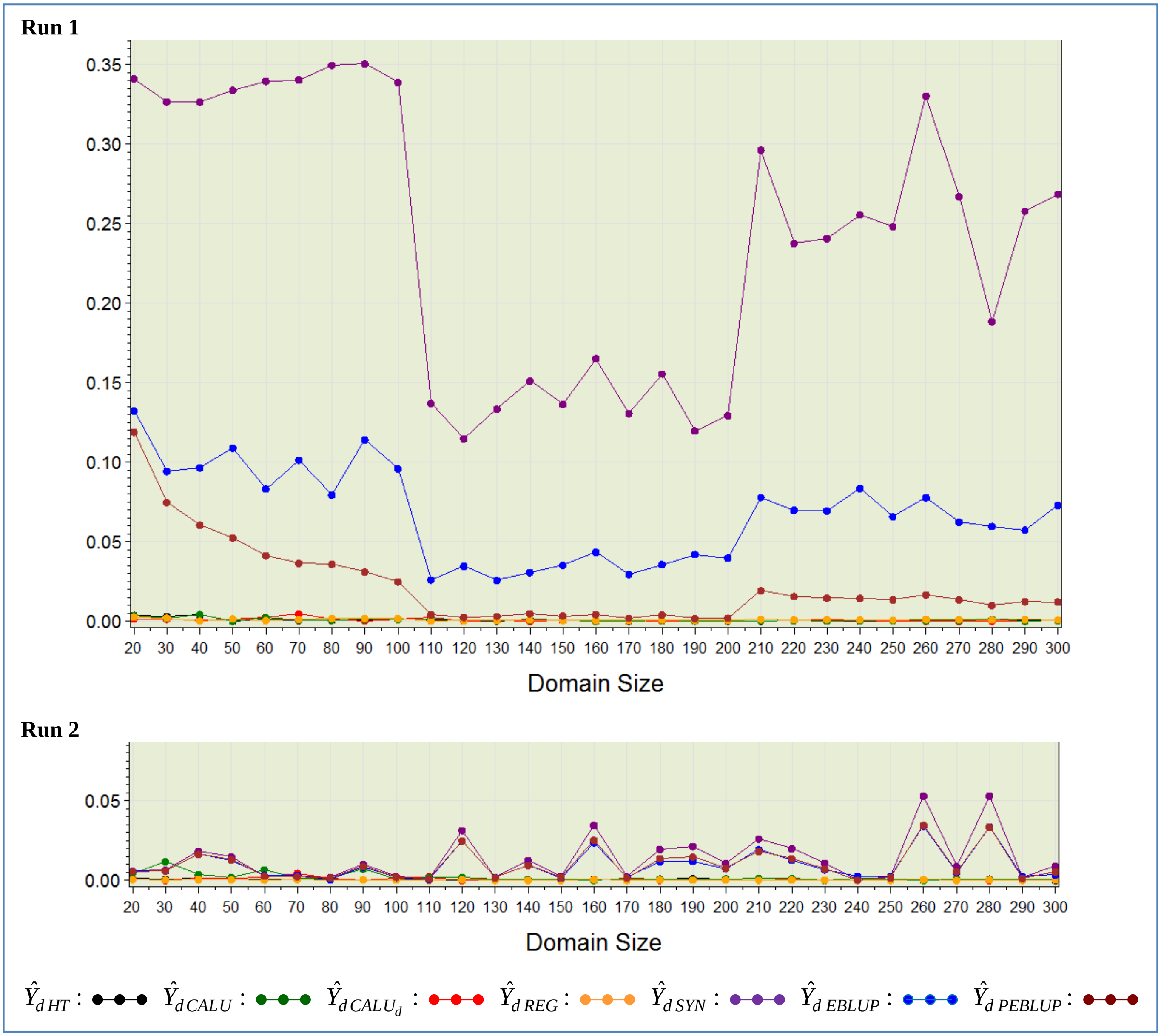

Figure 2: Plots of the Absolute Relative Bias of the Estimators (for sample size 464)

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 3 shows the coefficient of variation associated with the results in Figure 2. 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 mathematically equivalent in the two runs).

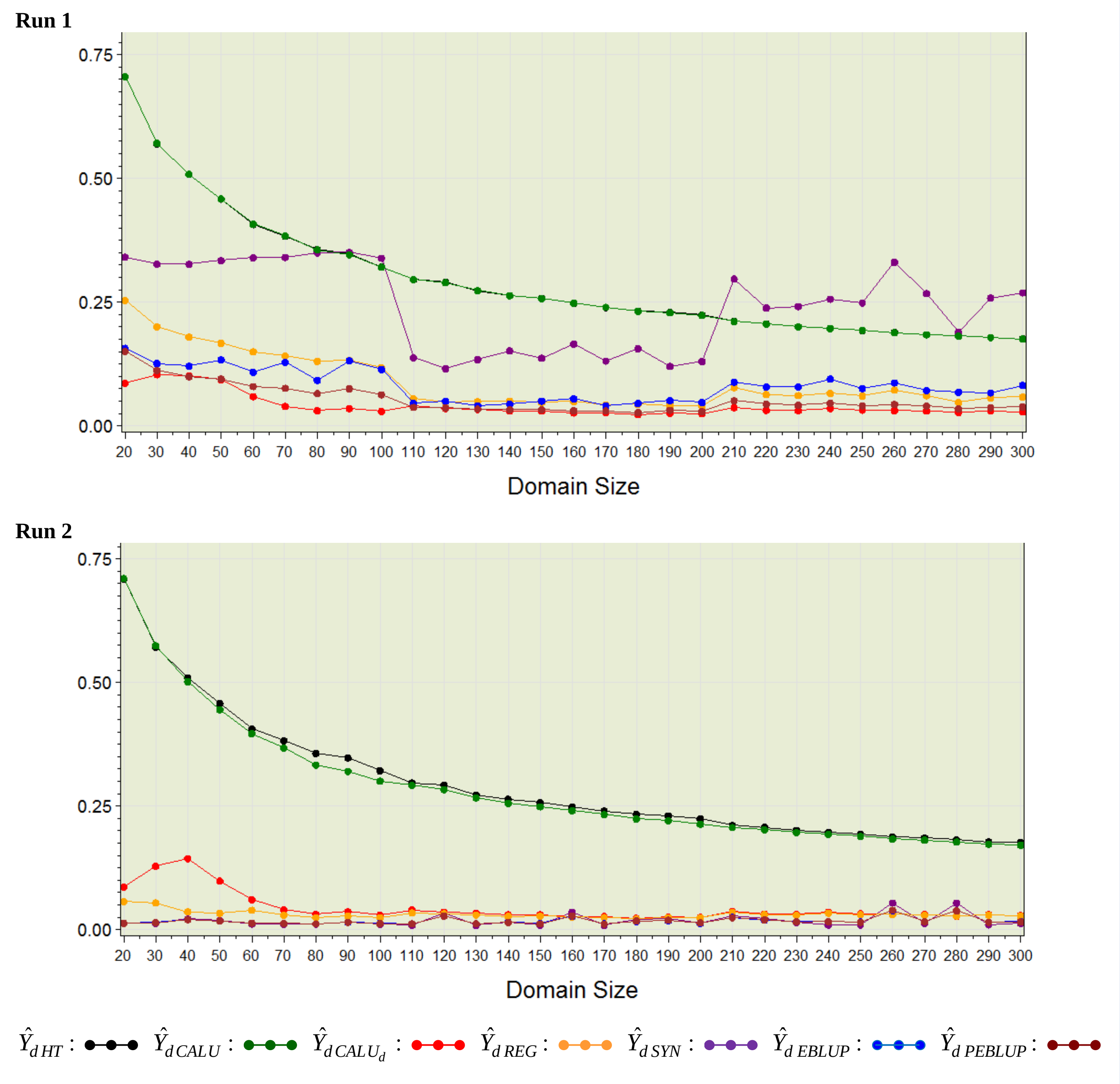

Figure 3: Plots of the Coefficient of Variation of the Estimators (for sample size 464)

This figure shows a graphical display of the results in Table 5. 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.

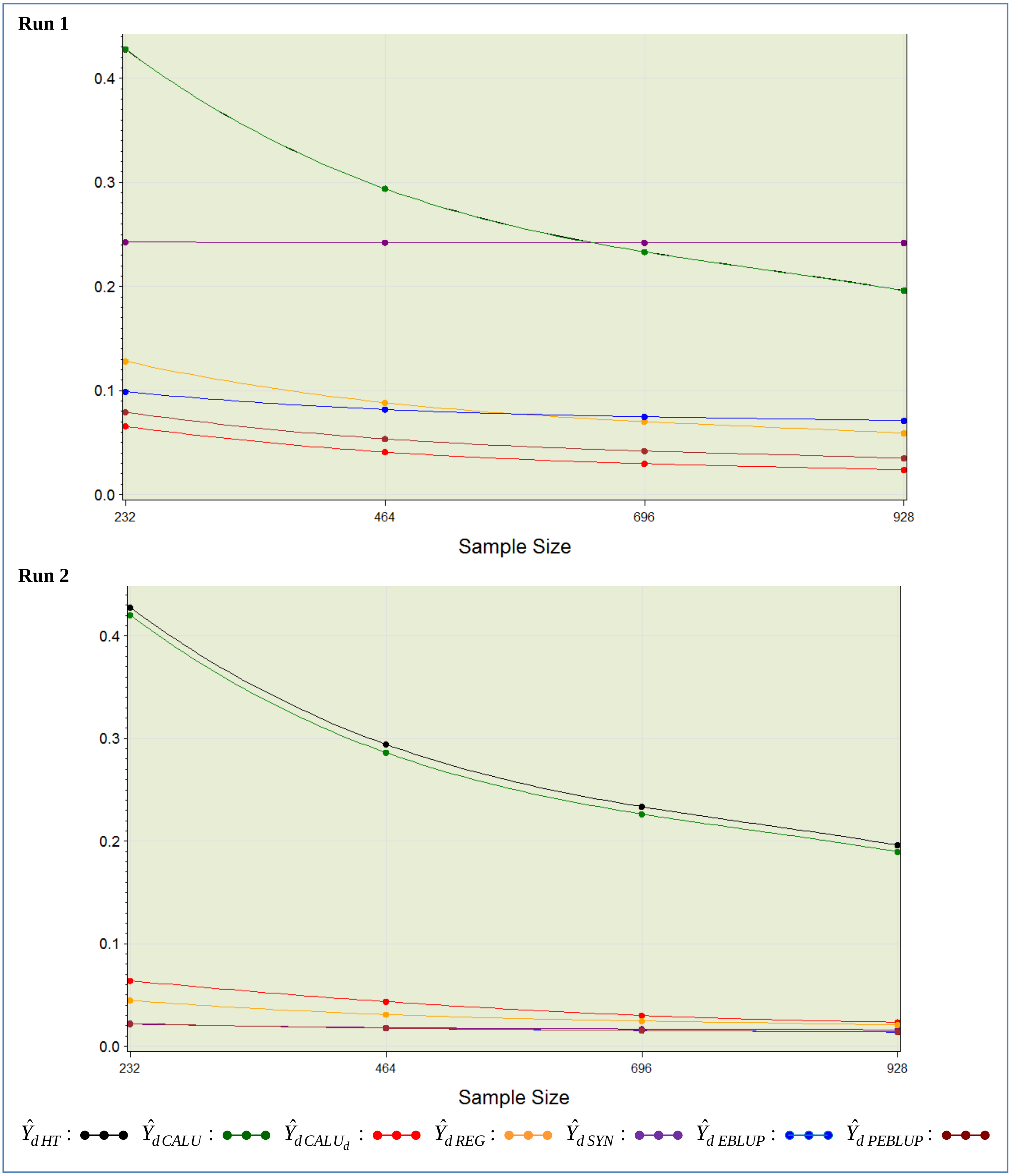

Figure 4: Plots of the Average Coefficient of Variation of the Estimators (by sample size)

This figure shows a graphical display of the results in Table 6. 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.

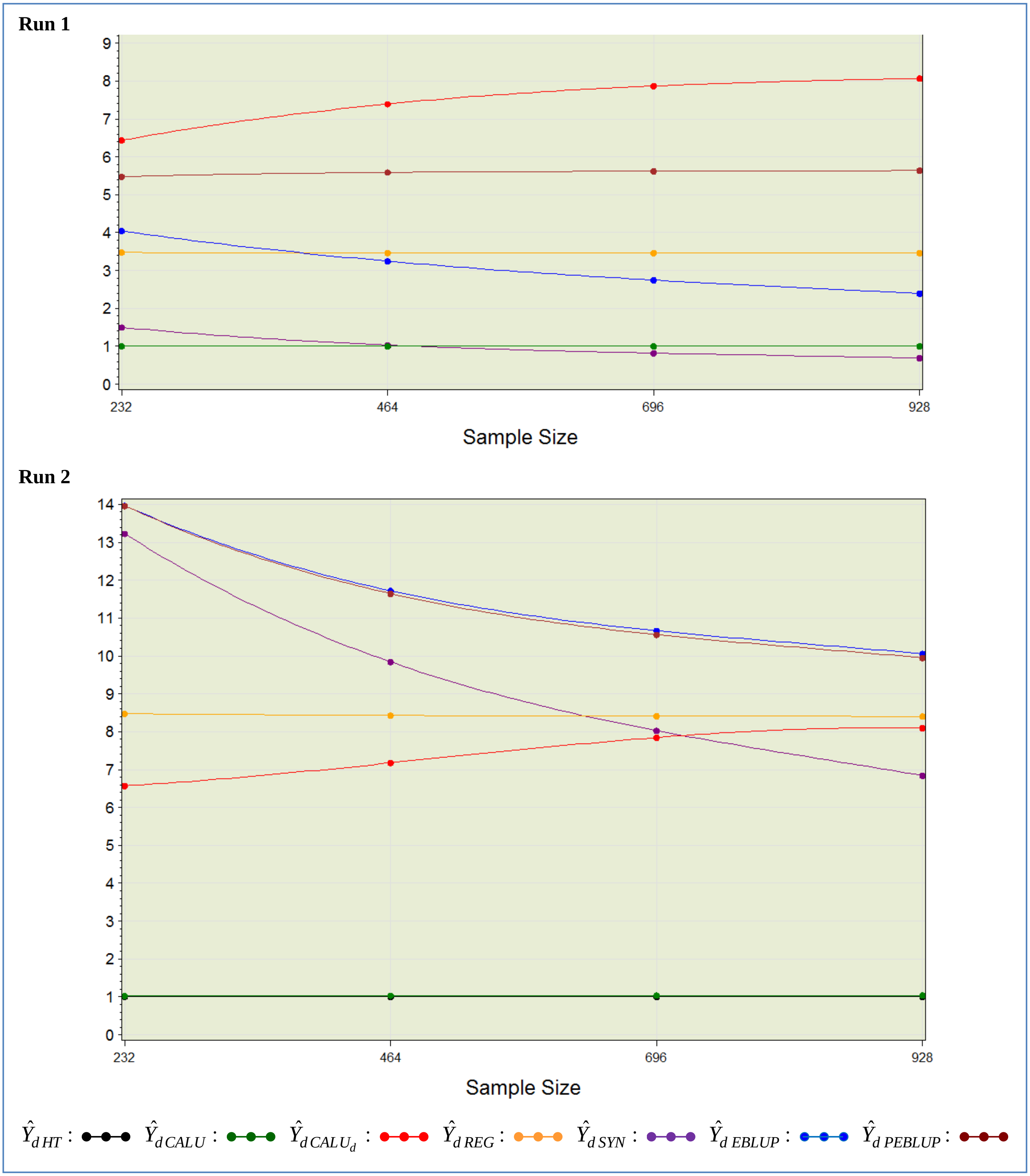

Figure 5: Plots of the Average Relative Efficiency of the Estimators (by sample size)

6 Conclusions

In summary, the comparison of the two small-area estimators reduces to the following conclusions. When we specify a 'correct' model, both estimators are equally efficient and they are better than any of the traditional estimators including $\hat{Y}_{CALU_d}$. When the specify an 'incorrect' model, $\hat{Y}_{PEBLUP}$ has higher efficiency than $\hat{Y}_{EBLUP}$. This is due to the use of the design weights to provide protection against model failure.

In those cases where we provide an 'incorrect' model, $\hat{Y}_{CALU_d}$ usually performs slightly better than the small-area estimators. The drawback with the use of $\hat{Y}_{CALU_d}$ is that it requires a sufficient number of units in each domain of interest. This is not always possible. It is also important to note that we tend to underestimate the variance of $\hat{Y}_{CALU_d}$ through the use of design-based variance estimation methods such as those used in generalized software such as GES. These problems are not present with the small-area estimators.

References

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.

Deville, J.C., and Särndal, C.E. (1992). Calibration estimation in survey sampling. *Journal* of *the American Statistical Association,* 87, 376-382.

Estevao, V., Hidiroglou, M.A., and You Y. (2014a). Methodology Software Library - Small-Area Estimation Fay-Herriot Area Level Model with EBLUP Estimation Methodology Specifications.

Estevao, V., Hidiroglou, M.A., and You Y. (2014b). Methodology Software Library - Small-Area Estimation Unit Level Model with EBLUP and Pseudo EBLUP Estimation Methodology Specifications.

Fay, R.E. and 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, 269-277.

Gonzalez, M.E. (1973). Use and Evaluation of Synthetic Estimates. *Proceedings of the Social Statistics Section,* American Statistical Association, pp. 33-36.

Rao, J.N.K. (2003). Small Area Estimation, Hoboken, New Jersey: John Wiley & Sons.

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, 496-504.

You, Y., and 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**.**