DOMAIN ESTIMATION USING LINEAR REGRESSION

M. Hidiroglou and Z. Patak
Zdenek Patak, 11-Q, R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada

Key words: Domain estimation, auxiliary data, conditional properties.

1. INTRODUCTION

One of the main objectives of a sample survey is to compute estimates of means and totals of a number of characteristics associated with the units of a finite population *U*. The data are often used for analytic studies or analyses of a survey. This usually involves the comparison of means and totals for subgroups of the population. Such subgroups are referred to as *domains of study*. Hartley's (1959) paper is one of the first attempts to unify the theory for domains. Hartley provided the theory for a number of sample designs where domain estimation was of interest. His paper mostly discussed estimators that did not make use of auxiliary information. He did, however, consider the case of the ratio estimator where population totals were known for the domains. The existence of multivariate auxiliary data raises a number of questions in the context of domain estimation. Some of those questions are as follows. What is the effect of having auxiliary information that is not known on a population basis for the given domain of interest? How do we compute valid variance estimates in the context of domain estimators that use auxiliary data? If more than one estimator is possible for point estimation and/or variance estimation, what criteria should be used to decide on how to chose the best estimator?

Durbin (1969) supported the use of conditional inference to do such comparisons. To quote him, he stated, "If the sample size is determined by a random mechanism and one happens to get a large sample, one knows perfectly well that the quantities of interest are measured more accurately than they would have been if the sample size had happened to be small. It seems self evident that one should use the information available on sample size in the interpretation of the result. To average over variations in sample size which might have occurred but did not occur, when in fact the sample size is exactly known, seems quite wrong from the standpoint of the analysis of the data actually observed". Holt and Smith (1979) favored conditional inference, and applied it to study the properties of the post-stratified estimator, given simple random sampling. Rao (1985) introduced the idea of "recognizable subsets" of the population to formalize the conditioning process. Recognizable subsets are defined *after* the sample has been drawn. In the context of domain estimation the number of units belonging to a particular domain is a random variable. Recognizable subsets in that context are those where the sample size is fixed within each domain. Comparison of the conditional statistical properties (i.e., bias, mean squared error) of the different estimators can then be based on these subsets. The conditioning process is that population totals are known for each domain. In the case of simple random sampling, the number of units in the population domain is assumed known.

The main purpose of this paper is to study the properties of a number of domain estimators of totals in the presence of auxiliary data. These properties will be established via conditioning on fixed sample sizes within each domain.

2. USE OF AUXILIARY INFORMATION IN DOMAIN ESTIMATION

Some notation is required to define the problem. Let the finite population $U = \{1,...,k,...,N\}$ be divided into *D* non-overlapping domains $U_1,...,U_d,...,U_D$. Let $Y = \sum_U y_k$ be the population total of a characteristic of interest "*y*". Assume that the sampling plan, *P(s)*, is an arbitrary one with first and second order inclusion probabilities π_k and $\pi_{k\ell}$. The resulting sample is denoted "*s*", and units in domain U_d that are part of *s* are denoted $s_d = U_d \cap s$. An estimator of the domain total $Y_d = \sum_{U_d} y_k$ that does not use auxiliary data is given by

$$\hat{Y}_{d,HT} = \sum_{s_d} w_k\, y_k = \sum_s w_k\, y_{dk},$$

where $w_k = \pi_k^{-1}$, y_{dk} is equal to y_k if $k \in s_d$ and 0 otherwise.

Auxiliary information in the form of a *p*-dimensional vector $\mathbf{x}$ may be available at different levels of aggregation. It may be known for each unit in the population, or for subsets $U_g \subseteq U$ $(g = 1,...,G)$ of the population *U* that may coincide with the domains U_d. We denote such totals $\mathbf{X}_g = \sum_{U_g} \mathbf{x}_k$, and they are estimated by $\hat{\mathbf{X}}_{g,HT} = \sum_{s_g} w_k\, \mathbf{x}_k$. New weights $\tilde{w}_k$ incorporating the auxiliary data can either be

constructed via calibration or linear regression procedures (LR). We chose the LR approach. In the case of *G* groups, the LR estimator is given by

$$\hat{Y}_{\ell r} = \hat{Y}_{HT} + \sum_{g=1}^{G} (\mathbf{X}_g - \hat{\mathbf{X}}_{g,HT})'\hat{\mathbf{B}}_g ,$$

where

$\hat{\mathbf{B}}_g = \left(\sum_{s_g} w_k \mathbf{x}_k \mathbf{x}_k' / c_k \right)^{-1} \sum_{s_g} w_k \mathbf{x}_k y_k / c_k$, and c_k are suitable positive constants.

The use of auxiliary data in the domain context offers a wide range of choices for various levels at which auxiliary totals are used and regression models are constructed. To simplify matters, we assume that g=1 (e.g.: a single group *U*), yielding the simple regression estimator $\hat{Y}_{\ell r} = \hat{Y}_{HT} + (\mathbf{X} - \hat{\mathbf{X}}_{HT})'\hat{\mathbf{B}}$. We consider six possibilities for estimating the domain population total Y_d. These possibilities differ based on whether we use the domain totals $\mathbf{X}_d$ or the population total $\mathbf{X}$, and whether we construct the regression estimator at the domain or at the population levels. The estimators are categorized into Horvitz-Thompson and "Hájek" types. We will elaborate on the difference between them in the next two subsections.

2.1. Horvitz-Thompson type estimators

Case 1

We assume that the auxiliary information is only available at the population level *U,* and that the regression model is estimated using the entire sample s. The dependent variable is y_{dk}, taking on the value y_k if *k* is in the domain and 0 otherwise, and $\mathbf{x}_k$ is the explanatory variable vector. The corresponding regression parameter is

$$\hat{\mathbf{B}}_1 = \left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \sum_s \frac{w_k \mathbf{x}_k y_{dk}}{c_k} ,$$

and the resulting estimator of the population total Y_d is

$$\hat{Y}_{d,\ell r_1} = \hat{Y}_{d,HT} + (\mathbf{X} - \hat{\mathbf{X}}_{HT})'\hat{\mathbf{B}}_1 . \qquad (2.1)$$

Case 2

If the auxiliary data totals are available at the domain level, $\mathbf{X}_d = \sum_{U_d} \mathbf{x}_k$, then two possible estimators of Y_d can be constructed depending on how the population regression ***B*** parameter is estimated. If the population parameter is estimated by using the observations within the sample s_d, then the resulting regression estimator is

$$\hat{Y}_{d,\ell r_2} = \hat{Y}_{d,HT} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HT})'\hat{\mathbf{B}}_2 , \qquad (2.2)$$

where $\hat{\mathbf{B}}_2 = \left(\sum_{s_d} \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \sum_{s_d} \frac{w_k \mathbf{x}_k y_k}{c_k}$.

Case 3

If the parameter is estimated using the entire available sample information, then the regression estimator is

$$\hat{Y}_{d,\ell r_3} = \hat{Y}_{d,HT} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HT})'\hat{\mathbf{B}}_3 , \qquad (2.3)$$

where $\hat{\mathbf{B}}_3 = \left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \sum_s \frac{w_k \mathbf{x}_k y_k}{c_k}$.

2.2. Hájek type estimators

Estimators (2.1)-(2.3) are of a Horvitz-Thompson nature. The "Hájek" versions of these cases are obtained by replacing $\hat{Y}_{d,HT}$, $\hat{\mathbf{X}}_{d,HT}$, and $\hat{\mathbf{X}}_{HT}$ by $\hat{Y}_{d,HA} = N_d \frac{\sum_{s_d} w_k y_k}{\sum_{s_d} w_k}$, $\hat{\mathbf{X}}_{d,HA} = N_d \frac{\sum_{s_d} w_k \mathbf{x_k}}{\sum_{s_d} w_k}$, and by $\hat{\mathbf{X}}_{HA} = N \frac{\sum_s w_k \mathbf{x_k}}{\sum_s w_k}$. The resulting estimators are:

Case 4

$$\tilde{\tilde{Y}}_{d,\ell r_1} = \hat{Y}_{d,HA} + (\mathbf{X} - \hat{\mathbf{X}}_{HA})'\hat{\mathbf{B}}_1 \qquad (2.4)$$

Case 5

$$\tilde{\tilde{Y}}_{d,\ell r_2} = \hat{Y}_{d,HA} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HA})'\hat{\mathbf{B}}_2 \qquad (2.5)$$

and

Case 6

$$\tilde{\tilde{Y}}_{d,\ell r_3} = \hat{Y}_{d,HA} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HA})'\hat{\mathbf{B}}_3 \qquad (2.6)$$

3. PROPERTIES OF THE DOMAIN ESTIMATORS

It is not clear which of the above estimators is the best in terms of the smallest bias and MSE. Such a comparison should be done conditionally. Since conditional inference in survey sampling can be stated in the case of simple random sampling, we limit our study to it.

The choice between these estimators will depend on the conditional bias and conditional mean squared error associated with each.

Estimators (2.1) - (2.6) may be expressed as:

$$\hat{Y}_{d,\ell r} = \sum_s w_k \, a_{dk} \, y_{dk} \qquad (2.7)$$

where a_{dk} is an adjustment factor that may or may not be domain dependent. Table 1 provides a summary of these factors. The adjustment factors for the Hájek analogues of $\hat{Y}_{d,\ell r_2}$ and $\hat{Y}_{d,\ell r_3}$ are obtained by replacing 1 with $N_d / \sum_{s_d} w_k$, and HT with HA in the definition of a_{dk} in Table 1. However, by virtue of $\hat{Y}_{d,HA}$ the a_{dk}'s for $\tilde{\tilde{Y}}_{d,\ell r_1}$ are now domain dependent and resemble those of $\hat{Y}_{d,\ell r_3}$ rather than $\hat{Y}_{d,\ell r_1}$.

Table 1: Adjustment Factors for Horvitz-Thompson Regression Estimators

Estimator	**Adjustment Factor** a_{dk}	**Residual** e_k
$\hat{Y}_{d,\ell r_1}$ *Domain Independent*	$1+(\mathbf{X}-\hat{\mathbf{X}}_{HT})'\left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1} \frac{\mathbf{x}_k}{c_k}, k \in s$	$\begin{cases} y_k - \mathbf{x}_k'\hat{\mathbf{B}}_1 & \text{if } k \in s_d \\ -\mathbf{x}_k'\hat{\mathbf{B}}_1 & \text{otherwise} \end{cases}$
$\hat{Y}_{d,\ell r_2}$ *Domain Dependent*	$\begin{cases} 1+(\mathbf{X}_d-\hat{\mathbf{X}}_{d,HT})'\left(\sum_{s_d} \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1} \frac{\mathbf{x}_k}{c_k}, k \in s_d \\ 0 \quad \text{otherwise} \end{cases}$	$\begin{cases} y_k - \mathbf{x}_k'\hat{\mathbf{B}}_2 & \text{if } k \in s_d \\ 0 & \text{otherwise} \end{cases}$
$\hat{Y}_{d,\ell r_3}$ *Domain Dependent*	$\begin{cases} 1+(\mathbf{X}_d-\hat{\mathbf{X}}_{d,HT})'\left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1} \frac{\mathbf{x}_k}{c_k}, k \in s_d \\ 0+(\mathbf{X}_d-\hat{\mathbf{X}}_{d,HT})'\left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1} \frac{\mathbf{x}_k}{c_k}, k \notin s_d \end{cases}$	$y_k - \mathbf{x}_k'\hat{\mathbf{B}}_3$

The corresponding variance expression for simple random sampling without replacement is given by

$$v(\hat{Y}_{d,\ell r}) = N^2 \frac{1-f}{n} \frac{\sum_s (a_{dk} e_k - \overline{a_d e})^2}{n-1}, \quad (2.8)$$

where a_{dk}'s are the appropriate adjustment factors and $e_k = y_{dk} - \mathbf{x}_{dk}'\hat{\mathbf{B}}_j$, (j = 1, 2, 3) are estimated residuals that depend on the regression estimator used, and $\overline{a_d e} = \frac{1}{n}\sum_{k=1}^{n} a_{dk} e_k$. We will provide examples of the required computations in the next section.

3.1 Conditional Bias

We examine the conditional bias of the estimators (2.1) to (2.6). As all these estimators are unconditionally unbiased, we compare their unconditional variance for the case where the error structure c_k is proportional to $\boldsymbol{\lambda}'\mathbf{x}_k$, where $\boldsymbol{\lambda}$ is an arbitrary vector. This last condition is satisfied when there is an intercept in the regression model linking the independent y_k variables to the $\mathbf{x}_k$ dependent variables, or when the error structure c_k is proportional either to one of the elements in the $\mathbf{x}_k$ vector or to a linear combination of them (i.e. $c_k = \boldsymbol{\lambda}'\mathbf{x}_k$).

We assume that the survey design is simple random sampling without replacement, as this greatly facilitates the task of obtaining conditional expectations. For a given sample domain s_d , let n_d be the realized sample size, and let $\mathbf{z}_k = (z_{k1},...,z_{kp})$. The following result can be used to evaluate the conditional bias of estimators (2.1) to (2.6).

Result 1: Given that $\mathbf{z}_k$ is a p-dimensional vector, and that $n_d \geq 1$, the conditional expectation of $\bar{\mathbf{z}}_s$ (the sample mean of the $\mathbf{z}_k$'s) is given by:

$$E(\bar{\mathbf{z}}_s | n_d) = \bar{\mathbf{z}}_U + \frac{w_d - W_d}{1-W_d}(\tilde{\mathbf{z}}_{U_d} - \bar{\mathbf{z}}_U) \qquad (3.1)$$

where: $\bar{\mathbf{z}}_s = n^{-1}\sum_s \mathbf{z}_k$, $\bar{\mathbf{z}}_U = N^{-1}\sum_U \mathbf{z}_k$;
$\tilde{\mathbf{z}}_{U_d} = N_d^{-1}\sum_{U_d} \mathbf{z}_k$; $w_d = n_d / n$; and $W_d = N_d / N$.

Remark 1: The proof of this result is obtained using conditional arguments. A consequence of the above result is that an estimator $\hat{\boldsymbol{\theta}}_s$ (say) will be nearly conditionally unbiased for its population parameter $\boldsymbol{\theta}_U$ (say) if either w_d is close to W_d , or if $\tilde{\boldsymbol{\theta}}_{s_{U_d}}$ (the population domain mean for U_d) is close to $\boldsymbol{\theta}_U$.

Result 2: If $\mathbf{z}_k$ is the domain variable y_{dk}, then we obtain from result 1 that $E(\hat{Y}_{d,HT}\,|\,n_d) = N\, w_d\, \tilde{y}_{U_d}$. The conditional bias of $\hat{Y}_{d,HT}$ is therefore:

$$\begin{aligned} Bias\,(\hat{Y}_{d,HT}\,|\,n_d) &= E\,(\hat{Y}_{d,HT}\,|\,n_d) - Y_d \\ &= N\,(w_d - W_d)\,\tilde{y}_{U_d} \end{aligned} \qquad (3.2)$$

Remark 2: The post-stratified analogue of $\hat{Y}_{d,HT}$, $\hat{Y}_{d,POST} = \dfrac{N_d}{n_d}\sum_s y_d$, is conditionally unbiased (Rao, 1985).

Result 3: Ignoring terms of $O(1/n)$, the approximate conditional bias of $\hat{Y}_{d,RAT} = \dfrac{X}{\hat{X}_{HT}}\hat{Y}_{d,HT}$ can then be shown to be:

$$N\,(w_d - W_d)\tilde{y}_{U_d}\,\frac{(\bar{x}_U - W_d\tilde{x}_{U_d})}{(\bar{x}_U - W_d\tilde{x}_{U_d}) + w_d(\tilde{x}_{U_d} - \bar{x}_U)}$$

If either $w_d \equiv W_d$ ($\equiv$ means close to) or $\bar{x}_U \equiv W_d\tilde{x}_{U_d}$ (implying that $U_d = U$), then this bias is approximately equal to 0. This conditional bias can be worse than the one associated with $\hat{Y}_{d,HT}$ if $\bar{x}_U > \tilde{x}_{U_d}$: hence, in that case it is better to use $\hat{Y}_{d,HT}$.

The above results can be shown using Result 1 and the Taylor series approximation method. Noting that the difference $\hat{Y}_{d,RAT} - Y_d$ can be written as $\hat{Y}_{d,RAT} - Y_d = \dfrac{X(\hat{Y}_{d,HT} - R_{1d}\,\hat{X})}{\hat{X}_{HT}}$ where $R_{1d} = Y_d / X$, the conditional bias of $\hat{Y}_{d,RAT}$ given n_d is

$$E\left[(\hat{Y}_{d,RAT} - Y_d)\,\middle|\,n_d\right] = X\;E\left[\left(\frac{\hat{Y}_{d,HT} - R_{1d}\,\hat{X}}{\hat{X}_{HT}}\,\middle|\,n_d\right)\right] \qquad (3.3)$$

The first order Taylor series expansion of $\hat{X}_{HT}^{-1}$ around $E(\hat{X}_{HT}\,|\,n_d)$ is

$$\frac{1}{\hat{X}_{HT}} = \frac{1}{E(\hat{X}_{HT}\,|\,n_d)}\left(1 - \frac{\hat{X}_{HT} - E(\hat{X}_{HT}\,|\,n_d)}{E(\hat{X}_{HT}\,|\,n_d)}\right), \qquad (3.4)$$

and from Result 1, we have that

$$E\,(\hat{X}_{HT}\,|\,n_d) = N\left[\bar{x}_U + \frac{w_d - W_d}{1 - W_d}(\tilde{x}_{U_d} - \bar{x}_{U_d})\right] \qquad (3.5)$$

We obtain the required result by substituting (3.4) and (3.5) into (3.3), and simplifying the algebra.

Remark 3: Using similar arguments, it can be shown that the conditional bias of the post-stratified ratio estimator $\hat{Y}_{d,POSTR} = \dfrac{X_d}{\hat{X}_{d,HT}}\hat{Y}_{d,HT}$ is negligible.

Result 4: The conditional biases of $\hat{Y}_{d,\ell r_j}$ (j = 1, 2, 3) can be obtained using similar arguments. For instance, it can be shown that the conditional bias of $\hat{Y}_{d,\ell r_1}$ is approximately given by:

$$\begin{aligned} Bias\,(\hat{Y}_{d,\ell r_1}\,|\,n_d) &= N\,(w_d - W_d)\,\tilde{y}_{U_d} + \\ &\left[X - E\,(\hat{X}_{HT}\,|\,n_d)'\right]\left[E\,(\mathbf{G}_S\,|\,n_d)\right]^{-1} E\,(\mathbf{H}_{s_d}\,|\,n_d), \end{aligned} \qquad (3.6)$$

where the conditional expectations of

$\mathbf{G}_s = \left(\sum_s \mathbf{x}_k\,\mathbf{x}_k'/c_k\right)$ and $\mathbf{H}_{s_d} = \left(\sum_s \mathbf{x}_k\,y_{dk}/c_k\right)$ can be obtained using Result 1.

The Hájek versions of $\hat{Y}_{d,\ell r_j}$, such as the post-stratified estimator, are approximately conditionally unbiased.

3.2 Conditional Mean Squared Error

We would like to show that the conditionally unbiased estimators have a smaller conditional mean squared error than their conditionally biased counterparts.

Remark 4: The conditional mean squared error of the Horvitz-Thompson estimator, $\hat{Y}_{d,HT} = \hat{Y}_d$, is uniformly greater than that of the post-stratified count estimator, $\hat{Y}_{d,POSTC} = \hat{Y}_d\,\dfrac{N_d}{\hat{N}_d}$. We consider two cases: (i) $n_d \geq \dfrac{n}{N}N_d$, and (ii) $n_d < \dfrac{n}{N}N_d$. When case (i) occurs, the result follows directly from the expressions of conditional variance. For case (ii) we have to show that $\left[n_d\,\dfrac{N}{n} - N_d\right] < \dfrac{1 - f_d}{n_d}\left[n_d\,\dfrac{N}{n} + N_d\right]$. Since this condition is always true, the result follows.

3.3 Unconditional Variance

The form of the population variance estimator given by expression (2.8) is unconditional. The unconditional population variances for the domain estimators with the auxiliary data available at the domain or population level can be compared directly.

Remark 5: The unconditional population variance of the ratio estimator, $\hat{Y}_{d,RAT} = \frac{X}{\hat{X}} \hat{Y}_{d,HT}$, is uniformly greater than that of the corresponding post-stratified counterpart, $\hat{Y}_{d,POSTR} = \frac{X_d}{\hat{X}_{d,HT}} \hat{Y}_{d,HT}$. The proof follows from showing that the regression parameter, which minimizes the residual sum of the squares within the domain, is $\hat{Y}_{d,HT} / \hat{X}_{d,HT}$.

Remark 6: The unconditional population variance of $\hat{Y}_{d(\ell r_2)}$ is smaller than the one for $\hat{Y}_{d(\ell r_1)}$, provided that N> p+ 1, where p+1 refers to the number of auxiliary variables (including an intercept term), SRSWOR sampling, and c_k is constant.

3.4. Specific domain estimators

We have selected eight estimators that we will classify as belonging to one of the preceding six cases. We shall examine the estimators' properties in terms of relative bias, relative mean squared error, relative variance, and coverage probability, along with their expected conditional behavior. This is done using a Monte Carlo simulation study.

Several estimators belong to the six cases. We target those that use a single auxiliary variable, be it categorical (count) or continuous (*x* variable), i.e., ratio type.

Case 1:
Auxiliary information is available at the population level. Linear regression modeling is performed at the level of the entire sample using the *y* variable as a domain variable. We consider *expansion* and *ratio* estimators defined as

$$\hat{Y}_{d,HT} = \hat{Y}_d \text{ and } \hat{Y}_{d,RAT} = \hat{Y}_d \frac{X}{\hat{X}}.$$

The residual terms used in the computation of variance are defined in Table 2. The adjustment factor is one for the expansion estimator. For the ratio estimator we have a choice of either $\frac{X}{\hat{X}_{HT}}$ or one. Corresponding estimated variances are denoted v_1 and v_2. Since $\frac{X}{\hat{X}_{HT}}$ tends to 1 as both the population and sample sizes tend to infinity, we would not expect the two choices of a_{dk} to result in markedly different variances.

Case 2:
Auxiliary data are available at the domain level. Linear regression is confined to the domain level as well. Two estimators, *post-stratified count* and *post-stratified ratio*, are studied. The corresponding estimated domain totals are

$$\hat{Y}_{d,POSTC} = \hat{Y}_d \frac{N_d}{\hat{N}_d} \text{ and } \hat{Y}_{d,POSTR} = \hat{Y}_d \frac{X_d}{\hat{X}_d}.$$

The residuals are once again defined in Table 2. Just as we have done for Case 1, we use two forms of a_{dk}'s in the variance expressions. The first one is defined in Table 1 ($\hat{Y}_{d,\ell r_2}$), while the other is reduced to one.

Case 3:
Auxiliary information is available at the population level. However, linear regression is performed using the entire sample. Two estimators, *alternate expansion* and *alternate ratio*, are investigated. The corresponding expressions for estimated domain totals are

$$\hat{Y}_{d,ALTE} = \hat{Y}_d + (N_d - \hat{N}_d)\bar{y}_s$$

and

$$\hat{Y}_{d,ALTR} = \hat{Y}_d + (X_d - \hat{X}_d)\frac{\hat{Y}}{\hat{X}}.$$

The form of the residual terms is given in Table 2. Variances are again computed using a_{dk}'s as defined in Table 1 ($\hat{Y}_{d,\ell r_3}$), or by replacing the a_{dk}'s with one.

Case 4:
This is the first of the Hájek type estimator classes. It mirrors Case 1; however, by virtue of the Hájek adjustment it is now domain dependent. The corresponding *x*-variable ratio estimator, *Hájek ratio*, of a domain total is.

$$\hat{Y}_{d,HAJR} = \hat{Y}_d \frac{N_d}{\hat{N}_d} + (X - \hat{X})\frac{\hat{Y}_d}{\hat{X}}.$$

In the case of count auxiliary information the estimator reduces to (2.2). The residuals for variance calculations are provided in Table 2.

Case 5:
In the special case of the ratio estimator, the Hájek version of (2.2) reduces to its Horvitz-Thompson counterpart. This means that Case 5 reduces to Case 2.

Case 6:

To study the properties of estimators in this class, we use the following candidate

$$\hat{Y}_{d,MODR} = \hat{Y}_d \frac{N_d}{\hat{N}_d} + (X_d - \hat{X}_d \frac{N_d}{\hat{N}_d}) \frac{\hat{Y}}{\hat{X}},$$

denoted the *modified alternate ratio*. Estimated variances are computed using residuals corresponding to those of the alternate ratio estimator.

Table 2. Definition of error terms.

Estimator	Error Term	
HT (Case 1)	$e_k = y_{dk} - \bar{y}_s(d)$	$\bar{y}_s(d) = \frac{\hat{Y}_{d,HT}}{N}$
RAT (Case 1)	$e_k = y_{dk} - \tilde{R}_d x_k$	$\tilde{R}_d = \frac{\hat{Y}_{d,HT}}{\hat{X}_{HT}}$
POSTC (Case 2)	$e_{dk} = y_{dk} - 1_{dk} \bar{y}_{s_d}$	$\bar{y}_{s_d} = \frac{\hat{Y}_{d,HT}}{\hat{N}_d}$
POSTR (Case 2)	$e_k = y_{dk} - \hat{R}_d x_{dk}$	$\hat{R}_d = \frac{\hat{Y}_{d,HT}}{\hat{X}_{d,HT}}$
ALT-E (Case 3)	$e_k = y_k - \bar{y}_s$	$\bar{y}_s = \frac{\hat{Y}_{HT}}{N}$
ALT-R (Case 3)	$e_k = y_k - \hat{R} x_k$	$\hat{R} = \frac{\hat{Y}_{HT}}{\hat{X}_{HT}}$
RAT-H (Case 4)	$e_k = y_{dk} - \tilde{R}_d x_k$	$\tilde{R}_d = \frac{\hat{Y}_{d,HT}}{\hat{X}_{HT}}$
MOD-R (Case 6)	$e_k = y_k - \hat{R} x_k$	$\hat{R} = \frac{\hat{Y}_{HT}}{\hat{X}_{HT}}$

Note: 1_{dk} is an indicator variable defined as 1 if $k \in s_d$ and 0 otherwise.

4. SIMULATION STUDY

To assess the conditional and unconditional properties of all the estimators defined in Section 3, a small simulation study was run. The population auxiliary variable was generated using a Γ-distribution to reflect the highly skewed nature of business populations, to which the conditional approach is uniquely applicable. Many social surveys would benefit from this methodology as well, especially where demographic domains are very small. Also, taking into account that most auxiliary information is well correlated with survey data, the population dependent variable was generated via a ratio model.

The auxiliary variable is generated using the gamma distribution $\Gamma(a,b)$, where $a = 3$ and $b = 16$. The dependent variable is also generated by a gamma distribution, $\Gamma(A,B)$. The parameters A and B are defined to satisfy $E(y_k) = \beta x_k$ and $V(y_k) = \sigma^2 x_k$. The correlation between X and Y is $\rho_{X,Y} = \frac{\beta b}{\sqrt{\sigma^2 b + \beta^2 b^2}}$. The population ratio $\beta = Y / X$ is specified by the user, and σ^2 is assigned a value to achieve the desired correlation. To allow for modeling small and large domains, a population of 1,000 observations was divided into two domains of 900 and 100 units. A common correlation value was used for both domains, 0.90 to represent a high correspondence between *x* and *y*, or then in decrements of 0.10 down to a correlation of 0.10 to examine the impact of a weakened $x - y$ relationship on the estimators under investigation. Slopes of $\beta_{d1} = 1.0$ and $\beta_{d2} = 3.0$ were used for large and small domains, with a sampling fraction of 0.25 for the entire sample resulting in the selection of 250 units for each iteration. A population ratio was assigned independently to each domain. For each combination of correlation coefficient and population ratio 100,000 iterations were executed to guarantee convergence.

To assess the unconditional properties of the estimators, several performance measures were computed. They were absolute relative bias (ARB), relative MSE efficiency (RMSE), and coverage rate (CR), each in turn given by

$$ARB_{d,\ell r} = \frac{100}{M} \left| \sum_{m=1}^{M} (\hat{Y}_{d,\ell r}^{(m)} - Y_d) / Y_d \right|, \text{ where } M \text{ is the number of replicates (100,000),}$$

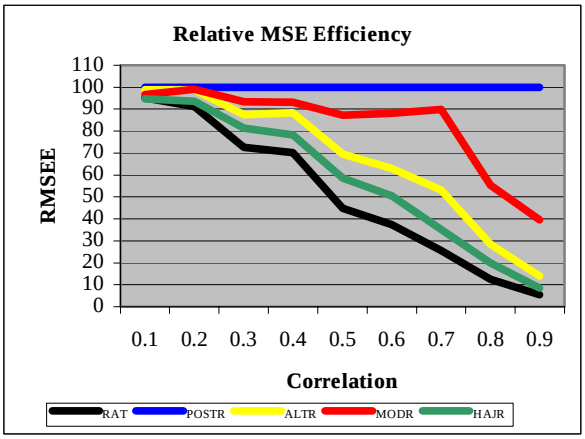
$$RMSE_{d,\ell r} = 100 \frac{\sum_{m=1}^{M} (\hat{Y}_{d,POSTR}^{(m)} - Y_d)^2}{\sum_{m=1}^{M} (\hat{Y}_{d,\ell r}^{(m)} - Y_d)^2},$$

and the coverage rate $CR_{d,\ell r}$ for a given estimator $\hat{Y}_{d,\ell r}$ is defined as the ratio of the number of times that the 95% confidence interval $\hat{Y}_{d,\ell r}^{(m)} \pm 1.96\sqrt{v(\hat{Y}_{d,\ell r}^{(m)})}$ contains the true population total to the number of replicates. Relative mean square error efficiency is given in terms of the post-stratified ratio estimator as it is virtually unbiased and serves as a good benchmark, against which the other estimators may be measured.

The two graphs below summarize the unconditional analysis in the context of small domains and gradually strengthening the relationship between *x* and *y* in terms of relative mean square error efficiency and coverage rates. We will comment only on the properties of the estimators in the case of large domains. The results for the absolute relative bias are not provided as it remains virtually unchanged as the relationship between x and y

weakens, since the estimators under investigation are unconditionally unbiased.

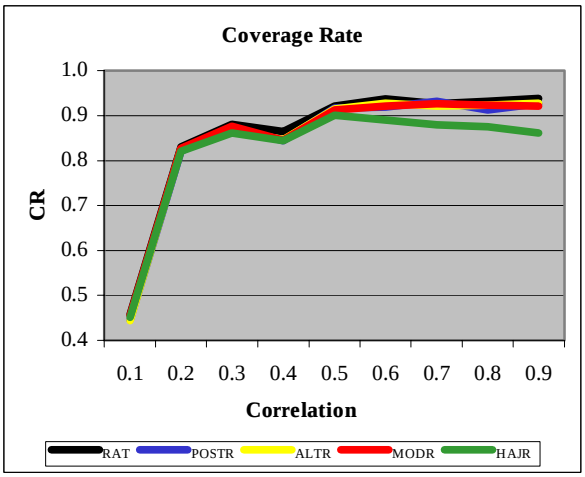
Figure 1. Unconditional relative MSE efficiency.



The relative MSE efficiency drops off substantially with the strengthening $x-y$ relationship. Part of this phenomenon can be attributed to the increasing dispersion of the dependent variable as we weaken its correlation with the auxiliary vector. This is further aggravated by the relatively small size of the domain in question.

The post-stratified ratio offers a substantial improvement over the other estimators in the face of strengthening correlation between x and y. In the case of large domains its dominance is less pronounced as the estimators are less influenced by small sample instability.

Figure 2. Unconditional coverage rates.



The coverage rates are similar across all the estimators with the modified alternate ratio performing slightly better than the rest when the $x-y$ relationship is strong. As it weakens, it exhibits the same coverage rate characteristics as its counterparts. For large domains the advantage of the modified alternate ratio estimator disappears.

Taking into account both the relative MSE efficiency and coverage rates, the post-stratified ratio and the modified alternate ratio estimators have the best properties, with the latter having a slight edge over the former in terms of coverage rates in the context of small domains.

The conditional relative bias and coverage rates of the estimators are summarized graphically below for a range of realized domain sample sizes in the context of small domains, as well as a strong and weak $x-y$ relationship.

Figure 3. Conditional relative bias for $\rho_{X,Y}=0.90$, $\beta_{d1}=1.0$, and $\beta_{d2}=3.0$.

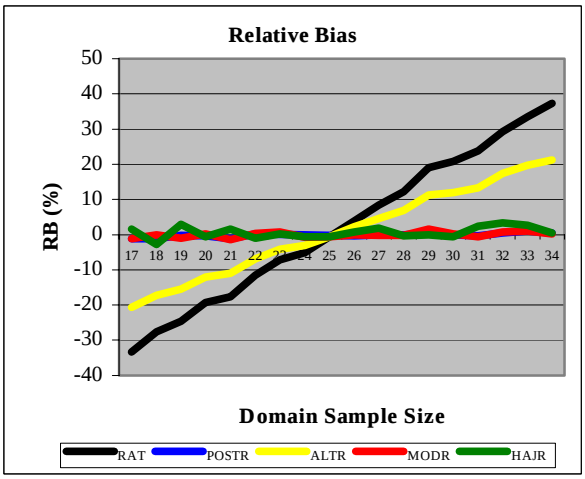


Figure 4a. Conditional coverage rates for $\rho_{X,Y}=0.90$, $\beta_{d1}=1.0$, and $\beta_{d2}=3.0$.

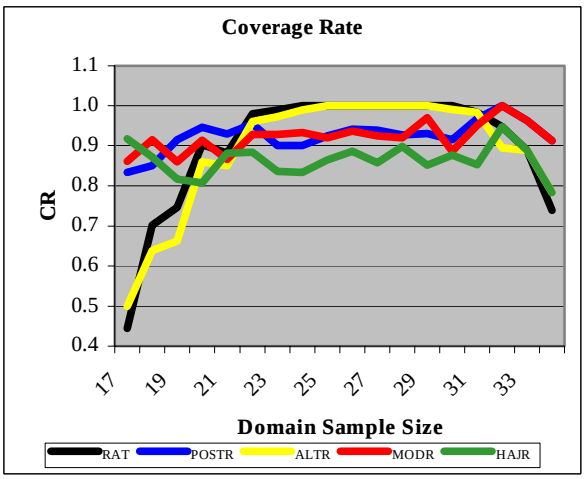
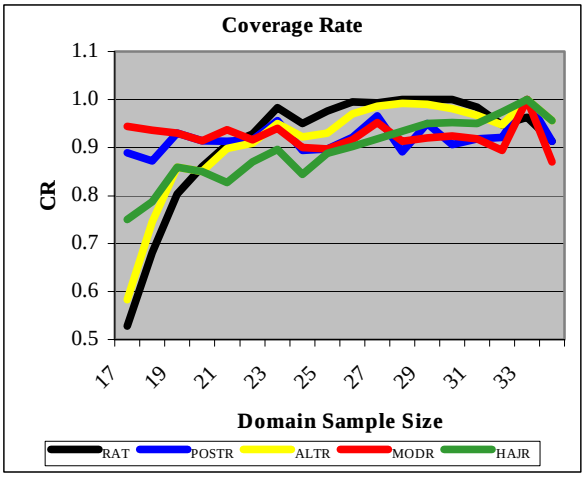


Figure 4b. Conditional coverage rates for $\rho_{X,Y} = 0.60$, $\beta_{d1} = 1.0$, and $\beta_{d2} = 3.0$.



5. CONCLUDING REMARKS

The simulation study identified two estimators, *post-stratified ratio* (POSTR) and *modified alternate ratio* (MODR), whose performance in terms of unconditional relative mean squared error efficiency and coverage rate, was superior to the other estimators under investigation. Within the confines of the simulation study, POSTR exhibited a weaker coverage rate than MODR due to its propensity to be influenced by small sample anomalies. This is true only if all domains within the sample behave similarly and one can gain strength from borrowing data across domains.

We can explain this by examining the difference in the definitions of POSTR and MODR. The former uses $\hat{Y}_d / \hat{X}_d$ while the latter does $\hat{Y} / \hat{X}$. MODR can be viewed as a semi-synthetic version of POSTR. It is much less affected by small sample related fluctuations, hence its superior coverage rate characteristics. Both outperform count based estimators over a wide range of correlation between *x* and *y*.

Another interesting observation worth noting is that calibrating to an auxiliary variable does not have a substantial impact on the magnitude of the mean squared error in the context of conditional analysis. In some cases calibration actually reduced the nominal coverage rates.

REFERENCES

Durbin, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling* (Eds. N.L. Johnson and H. Smith), New York: Wiley, Interscience.

Hartley, H.O. (1959). *Analytic Studies of Survey Data*. Instituto di Statistica, Rome, Volume in honor of Corrado Gini.

Holt, D. and Smith, T.M.F. (1979). *Post-stratification*. Journal of the Royal Statistical Society, Sec. A, 142, 33-46.

Rao, J.N.K. (1985). *Conditional Inferences in Survey Sampling*. Survey Methodology, 15-32.