Journal of Official Statistics, Vol. 20, No. 4, 2004, pp. 645–669

Borrowing Strength Is Not the Best Technique Within a Wide Class of Design-Consistent Domain Estimators

Victor M. Estevao[1] *and Carl-Erik Särndal*[2]

Estimation for subpopulations, or domains, is an important objective in most surveys, especially in large surveys conducted by national statistical agencies. These agencies practice design-based domain estimation whenever possible, that is, whenever the sample size is sufficient and auxiliary information is available. The precision, as measured by the design-based variance, is a function of these factors. Insufficient precision - leading to a suppression of estimates - is more likely to happen for minor domains than for major domains. Our starting point is a statement of the auxiliary information available for a survey. Strong information provides the material for precise domain estimates. We form a class of domain estimators based on the given auxiliary information. It includes regression fit estimators as well as calibration estimators, direct as well as indirect estimators. Direct estimators use only y-values from inside the domain itself. Indirect estimators borrow strength by incorporating external y-values thought to be "related." Borrowing strength is the cornerstone of small area estimation, a research tradition that is model-dependent, nondesign-based, and not examined in this article. The concept of borrowing strength is highly useful in that theory. However, since design-based domain estimation is extensively practiced, we are led to the question: What can borrowing strength do for design-based domain estimation? The answer is that borrowing strength is unfruitful in the design-based tradition. We find that for a fixed set of auxiliary information, the minimum asymptotic design-based variance is obtained with a direct estimator, derived by calibration rather than by regression fitting.

Key words: Design-based inference; very nearly design unbiased estimation; calibration; calibrated weights; regression fit; regression residuals.

1. Introduction

It is standard practice in a national statistical agency to provide estimates for the finite population of interest as well as for a number or subpopulations, called *domains* or *domains of study*. This activity relies on a research tradition known as design-based domain estimation. The inference is design-based, or randomization theory based. This inference perspective dominates in most national statistical agencies because of its objectivity and freedom from model assumptions. Hence national statistical agencies generally use design-based domain estimation. This is possible when the realized domain sample size and the available auxiliary information are sufficient to deliver an acceptable

[1] Senior Statistician, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6 . E-mail: victor.estevao@statcan.ca
[2] Professor, Consultant, Ottawa, Canada . E-mail: carl.sarndal@rogers.com
Acknowledgment: The authors sincerely thank the Associate Editor and five anonymous referees for their thorough reviews and comments. Their suggestions and insights helped significantly improve the presentation in the article.

design-based precision. Failing that, the agency may refrain from publishing an estimate or resort to model-dependent estimation.

Statistical agencies in several countries have developed, and continue to develop, software programs for design-based domain estimation. These programs compute design-consistent estimates for domains, and the corresponding design-based measures of precision. Examples with this general objective - although details may differ - include POULPE and CALMAR produced at INSEE in France, Bascula 4.0 created by Statistics Netherlands, CLAN97 built by Statistics Sweden, and the Generalized Estimation System (GES) developed by Statistics Canada. Their methodologies are described respectively in Caron, Deville, and Sautory (1998), Nieuwenbroek and Boonstra (2002), Andersson and Nordberg (1998), and Estevao, Hidiroglou, and Särndal (1995). The design-based domain estimation results in this article have a bearing on the further development of software in national statistical agencies.

Design-based inferences hold independently of the form of the sampled population, assuming there is no nonresponse. The estimates are viable as long as the domain sample size and the auxiliary information are sufficient. Failing this, the design-based estimates start to deteriorate. The conditions for design-based domain estimation are favourable in a number of European countries, because existing registers provide an excellent source of auxiliary information.

The article is arranged as follows. Section 2 discusses various aspects of domain estimation. Section 3 presents the background and the objectives for the article, which is concerned entirely with design-based domain estimation. The statement of auxiliary information is particularly important. Sections 4 and 5 present two approaches for constructing a design-based domain estimator from a fixed set of auxiliary information. These approaches are calibration and regression fitting. In Section 6, we define a wide class of estimators that encompasses both approaches. We conclude in Sections 7 to 9 that a direct estimator created by calibration has minimum asymptotic design-based variance, for the given auxiliary information. Any estimator in the class that attempts to borrow strength is less precise. Section 11 reports a Monte Carlo simulation. Its results are in agreement with the theoretical findings in Sections 7 to 9.

2. Terminology and Discussion

An estimator for a domain is commonly called *direct* if it only uses values y_k of the variable of interest for those units k that belong to the domain itself. An estimator is *indirect* if it also uses values y_k for units k outside the domain. The objective of national statistical agencies is to produce the best possible estimates for required domains given a specified accuracy and cost. This is generally done by producing design-based estimates as far as possible. Marker (2001) formulated this objective as follows: "While it is always possible to produce indirect, model-dependent, estimates for small areas, it is desirable to produce direct estimators where possible," and he notes that there are devices to meet this objective. Stratification, over-sampling and dual-frame estimation may be used to stretch the applicability of direct estimates. Another avenue is a systematic search for and use of multivariate auxiliary information at the estimation stage, as in the calibration method recommended in this article.

The *classification of domains by size* plays an important role. Purcell and Kish (1979) suggest a classification based on the relative size of domains. They distinguish *major*, *minor*, *mini*, and *rare* domains. Here we shall use the terms *major domain* and *minor domain*. There is no need to draw a firm line between the two categories. We can take a size of 10% of the population as a rough dividing line. The same minor domain may be successfully estimated with design-based domain estimation methods in one survey, while in another it may not.

A domain accounting for say 5% of the population is minor but in a survey designed to have a global sample size of $n = 30,000$ the expected domain sample size is as large as 1,500 under Simple Random Sampling Without Replacement (SRS). Depending on the auxiliary information available to strengthen the estimates, this may be sufficient for design-based estimation.

The asymptotic design-based domain estimation theory in this article is predicated by a total sample size n tending to infinity and a *bias ratio* (bias divided by standard deviation) of the domain estimator that approaches zero with order $O(n^{-1/2})$. In the words of one of the referees of the article, these estimators can be described as *very nearly design unbiased*. Under finite conditions, this theory still works as long as the domain sample size is not extremely small. For example, the results of our simulations show that the theory works well when applied to a domain as small as 10% of the population and with an expected domain sample size around 150.

In the Nordic countries, the existence of excellent registers makes it possible to make design-based estimates even for minor domains in the population of individuals. One makes sure that the sample selection and/or the use of auxiliary information will produce an acceptable precision with design-based domain estimation methods. However, in some other survey in the same country, that same small domain may end up having so few sampled units that the publication of design-based estimates is suppressed. If estimates are to be produced at all, they require small area estimation techniques. Ghosh and Rao (1994) note: "The terms small area and local area are commonly used to denote a small geographical area . . . They may also describe a "small domain," i.e., a small subpopulation . . . The usual direct survey estimates for a small area, based on data only from the sample units in the area, are likely to yield unacceptably large standard errors due to the unduly small size of the sample in the area."

The design-based direct estimates developed in this article are more sophisticated than the "usual direct estimates." Still, if the sample size in the domain is very small, the extremely large variance of the design-based domain estimator is likely to cause an erratic estimate not fit for use. To resolve this dilemma, model-dependent small area estimation has been developed over the last 30 years. Small area estimates usually have much smaller variance, thus are not prone to be erratic, but they have an unknown bias. Driving forces behind small area estimation are the desire (i) to provide alternatives when the design-based domain estimates do not meet standards in regard to precision and fitness for publication, and (ii) to meet the growing demand for estimates for small parts of a population, for example, smaller regions of a country. Small area estimation goes as far as producing estimates for a domain containing no observed values at all.

Small area estimation stands in contrast to the design-based domain estimation, not in its general objective, which, in both traditions, is to produce useful domain estimates, but

in the inference perspective. It relies on *model-dependent* inference. The extent to which small area estimates can be trusted hinges on the validity of the model assumptions. This is carefully noted in influential work on small area estimation.

The concept of *borrowing strength* is the cornerstone of small area estimation. In their review of small area estimation, Ghosh and Rao (1994) state: "Several powerful statistical methods with sound theoretical foundation have emerged for the analysis of local area data. Such methods "borrow strength" from related or similar small areas through explicit or implicit models that connect the small areas via supplementary data (e.g., census and administrative records)." For areas with insufficient sample size, Rao (1999) notes that "in making estimates for such small areas it is necessary to "borrow strength" from related areas to form "indirect" estimators that increase the effective sample size and thus increase the precision."

An attempt at borrowing strength occurs whenever an indirect estimator is used. One tries to overcome the insufficient design-based accuracy by incorporating y-values coming not from the domain itself but from outside, and assumed to be related. An attempt at borrowing strength can be deemed successful if the mean squared error (MSE) is smaller than that of a direct estimator. Borrowing strength amounts to a pooling of data. Borrowing strength and data pooling appeal to intuitive statistical instincts: A greater base of similar or related data should enhance the prospects of a reliable estimation for a domain.

Nevertheless, design-based domain estimation is widely practiced, in particular by national statistical agencies. This raises the question: What is the role of borrowing strength in design-based domain estimation? We prove that within the wide class of design-consistent estimators presented in Section 6, borrowing strength is not a fruitful concept. Some estimators in this class are constructed by regression fitting at different levels, others by calibration. Some are direct estimators. Others are indirect, that is, they attempt to borrow strength. Their common denominator is the fixed auxiliary information available for the survey. We conclude that, for the given auxiliary information, any estimator in the class cannot have smaller asymptotic design-based variance than the best direct estimator.

It is somewhat of a paradox to find that borrowing strength does not pay off for design-based domain estimation. That a traditional concept should be productive in one theory (model-dependent sampling theory) but not so in another (design-based sampling theory) is not an isolated occurrence. The concept of maximum likelihood is another example, as we note in the discussion in Section 12. One theory is not "better" than the other. Both are viable, but under different conditions.

3. Notation, Definitions, and Statement of Objective

We denote the finite population as $U = \{1, \ldots, k, \ldots, N\}$. The variable of interest is y and its value for unit k is y_k. One parameter of interest is the population total of y, denoted by $y_+ = \Sigma_U y_k$. The subscript "+" of a variable denotes the summation over $k \in U$. The notation Y is often used for this purpose, but this article requires sums for a variety of variables, so the subscript notation "+" is more efficient.

Let $U_d \subseteq U$ denote a domain of U. We need the domain specific y-variable, y_d, whose value for unit k is given by $y_{dk} = \delta_{dk} y_k$, where δ_{dk} is the domain identifier defined as $\delta_{dk} = 1$ for $k \in U_d$ and $\delta_{dk} = 0$ for $k \notin U_d$. Consequently, $y_{dk} = y_k$ for $k \in U_d$ and $y_{dk} = 0$ for $k \notin U_d$. The main parameter of interest in this article is the domain total of the variable y, $\Sigma_{U_d} y_k$, which is the population total of the variable y_d. That is, $y_{d+} = \Sigma_{U_d} y_k = \Sigma_U y_{dk}$.

A probability sample s is drawn from U with a given sampling design. The known nonzero inclusion probability for unit k is $\pi_k = P(k \in s)$ and the sampling weight of unit k is $a_k = 1/\pi_k$. The variable of interest is observed for all sample units, so the available y-data are $\{y_k : k \in s\}$. We assume no nonresponse. For every $k \in s$, we also observe membership in the domain. The unbiased, but often not very efficient Horvitz-Thompson (HT) estimator of y_{d+} is $\Sigma_s a_k y_{dk}$, which we denote by $y_{d\oplus\pi} = \Sigma_s a_k y_{dk}$. Our notation for sample weighted sums uses the following principle: the index $\oplus$ indicates a weighted sum over the units of the sample and the index π indicates sample weighting with $a_k = 1/\pi_k$. The first index, in this case d, identifies the variable, y_d, whose observed values y_{dk} are weighted and summed. The HT estimator is design unbiased for its population analogue, so $\mathrm{E}(y_{d\oplus\pi}) = y_{d+}$. The expected value operation removes the index π and the circle around $+$; the result is the unweighted population sum $y_{d+} = \Sigma_U y_{dk}$.

Domain estimation encounters two practical problems: (i) the sampling frame lists the population units but fails to identify the units belonging to the domain of interest; (ii) when the domain is small, the realized domain sample may be inadequate to meet the precision requirements. Here (i) implies that the size and other features of the domain are unknown. If the domain can be identified from the frame, action can be taken to obtain adequate sample size and/or auxiliary information for the domain and thereby an adequate precision for design-based domain estimates. Included in (i) is the difficulty that the domain code in the frame may be erroneous for some units, so the actual domain membership for these units is known only after they have been observed. Such classification errors are frequent in business surveys. In this article we deal with (i) and do not address (ii). We assume the domain sample size is not excessively small, and use design-based inference. The extensive literature on small area estimation addresses (ii) by model-dependent inference.

The use of auxiliary information is essential for efficient estimation. It consists of information on the variables that make up the vector $\mathbf{x}$ of dimension $J \geq 1$. Its value for unit k is denoted by $\mathbf{x}_k$. In this article, auxiliary information is viewed as consisting of two components: *Knowledge of aggregated values* $\mathbf{x}_k$ for one or more population groups called control groups, calibration groups or C-groups and *knowledge of individual values* $\mathbf{x}_k$ for the sampled units $k \in s$. The C-groups define the *C-level*. Our general notation for a calibration group is U_C. We need the C-group indicator, defined by $\delta_{Ck} = 1$ for $k \in U_C$ and $\delta_{Ck} = 0$ for $k \notin U_C$. For all $k \in U$, we define $\mathbf{x}_{Ck} = \delta_{Ck} \mathbf{x}_k$.

The *auxiliary information* about U_C has the following two components:

i) The auxiliary vector total $\mathbf{x}_{C+} = \Sigma_U \mathbf{x}_{Ck} = \Sigma_{U_C} \mathbf{x}_k$ is known.
ii) For every $k \in s$, the vector value $\mathbf{x}_k$ and membership or not of k in U_C are known.

Auxiliary information can come from different sources: the survey itself, a census, administrative registers, or a matching of such registers. The conditions (i) and (ii) are

present in two important practical circumstances: (a) The auxiliary vector total is "imported" from a reliable source unrelated to the survey itself; (b) There exists a population list, $k = 1, 2, \ldots, N$, with an $\mathbf{x}_k$-vector attached to every unit k, and this list serves as a list frame for drawing the sample s.

Consider first the case of an *imported total,* $\mathbf{x}_{C+}$. It must relate to the same concept as $\mathbf{x}_k$ in (ii). For example, $\mathbf{x}_{C+}$ must not be an out-of-date, erroneous total for the vector $\mathbf{x}_k$. In a survey on individuals, the C-groups may be defined by a crossing of regions with other categories. This happens in important surveys in North America. For example, the Canadian Labour Force Survey imports population figures (which are accurate census projections) for C-groups based on age category by sex by region (within a Canadian province). For each of these groups, we know the number of individuals from the census projection. The implied definition of $\mathbf{x}_k$ is $\mathbf{x}_k = 1$ for all individuals k. For group U_C, the auxiliary total is the group size N_C. But $\mathbf{x}_k$ need not be that elementary. If the totals in question can be imported, $\mathbf{x}_k$ may contain continuous as well as categorical variables.

Consider the case of a *population list* providing the information (i) and (ii), as is typical in surveys of individuals and households in several European countries, notably in the Nordic countries. The Register of the Total Population contains information on all persons k in the population (the frame) U. Consider the vector $\mathbf{x}_k$ containing the variables "years of education" and "salary", and consider a broad occupation category U_C. For every individual $k \in U$, and hence for every $k \in s$, we know both the vector value $\mathbf{x}_k$ and membership or not in U_C. Requirement (ii) is met. Suppose the objective is to estimate y_{d+} for a more narrow occupation category, U_d, contained in U_C but such that membership in U_d is not recorded on the frame. At the design stage we know which units $k \in U$ are in U_C, but not which ones are in U_d. Had this latter information been available, we could have designed the survey with U_d as a stratum with an adequate sample size, or as a C-group with a known total of a strong auxiliary vector. But we do not have access to $\mathbf{x}_{d+} = \Sigma_{U_d}\mathbf{x}_k$. However, by summing the $\mathbf{x}_k$ on the frame we compute the higher level total $\mathbf{x}_{C+} = \Sigma_{U_C}\mathbf{x}_k$, so requirement (i) is met. This information is still valuable in estimating for U_d. Ideally, we would like $U_C = U_d$, as Estevao and Särndal (1999) note, but we do not usually have auxiliary information at this level. In practice, we must often be content with information at a level above the domain.

In many surveys, the objective is to estimate the totals y_{d+} of a set of domains U_d, $d = 1, \ldots, D$. These may form a partition of the population U, as is often the case when the domains are for example the regions of a country. This partitioning into domains U_d is called the *d-level*. We focus, however, on one particular domain, U_d.

In Sections 4 to 9 and 11, we examine the case where U_C is a C-group containing the domain of interest U_d, so that $U_d \subseteq U_C \subseteq U$. In Section 10, we cover the case where U_d intersects several calibration groups.

Consider the case $U_d \subseteq U_C \subseteq U$. We know the C-group $\mathbf{x}$-total $\mathbf{x}_{C+} = \Sigma_U \mathbf{x}_{Ck}$. By contrast, the domain $\mathbf{x}$-total $\mathbf{x}_{d+} = \Sigma_U \delta_{dk}\mathbf{x}_k$ is unknown unless $U_C = U_d$. Special cases to be examined include: (i) The domain U_d is itself a C-group, below the level of the entire population: $U_d = U_C \subset U$; (ii) The whole population is a C-group, above the level of the domain: $U_d \subset U_C = U$. The individual value $\mathbf{x}_k$ is known for all sampled units, $k \in s$, as is membership or not of k in U_C. That is, we know $\mathbf{x}_k$ for every $k \in s_C = s \cap U_C$ (and for

every $k \in s_d = s \cap U_d$). We can form the HT estimator of the known total $\mathbf{x}_{C+}$ as $\mathbf{x}_{C\oplus\pi} = \Sigma_s a_k \mathbf{x}_{Ck} = \Sigma_{s_C} a_k \mathbf{x}_k$. It is unbiased since $\mathrm{E}(\mathbf{x}_{C\oplus\pi}) = \mathbf{x}_{C+}$.

We formulate the estimation problem as follows: For the domain of interest U_d such that $U_d \subseteq U_C \subseteq U$, we seek to estimate the unknown domain y-total, $y_{d+} = \Sigma_{U_d} y_k = \Sigma_U y_{dk}$. Available for this purpose are the data $\{(\mathbf{x}_k, y_k) : k \in s\}$, the C-total $\mathbf{x}_{C+}$, and its unbiased estimate $\mathbf{x}_{C\oplus\pi}$. What is the best use of this information for estimating y_{d+}? The answer is given in Section 8. First, we outline two reasonable approaches, the calibration approach (Section 4) and the regression fit approach (Section 5). Both approaches can be carried out in a variety of ways. For the same auxiliary information, we can thus create a variety of estimators. Regression fitting leads to generalized regression (GREG) estimators. Calibration, as presented for example in Deville and Särndal (1992), is also a well-known technique. In Section 6 we show that the calibration and the regression fitting approaches can be viewed as part of a more general class of estimators. The optimal estimator in that class is a calibration estimator.

4. The Calibrated Weights Approach

The calibrated weights (CALWEIGHT) approach relies on a system of calibrated weights, $w_k = a_k g_k$ for $k \in s$, computed with the given auxiliary information (i) and (ii) in Section 3. We apply these weights to the domain variable $y_{dk} = \delta_{dk} y_k$. This technique is used in CLAN97 and in GES. The resulting estimator is

$$\hat{y}_{d+} = \sum\nolimits_s w_k y_{dk} \tag{4.1}$$

where $w_k = a_k g_k$ and

$$g_k = 1 + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}_{Ck}^T \right)^{-1} \mathbf{z}_k$$

where $\mathbf{x}_{Ck} = \delta_{Ck} \mathbf{x}_k$, $\mathbf{x}_{C+} = \Sigma_U \mathbf{x}_{Ck}$ is the known C-total, $\mathbf{x}_{C\oplus\pi}$ is the corresponding HT estimator, and $\mathbf{z}_k$ is a J-vector satisfying the following conditions: $\mathbf{z}_k$ can have any value, including $\mathbf{0}$, as long as it is not $\mathbf{0}$ for all $k \in s$ and $(\Sigma_s a_k \mathbf{z}_k \mathbf{x}_{Ck}^T)$ is nonsingular. The use of an *instrument vector* $\mathbf{z}_k$ for the purposes of calibration is discussed for example in Estevao and Särndal (2000) and Deville (2002). For any such $\mathbf{z}_k$, the weights w_k are calibrated to the C-level. That is, $\Sigma_{s_C} w_k \mathbf{x}_k = \Sigma_s w_k \mathbf{x}_{Ck} = \mathbf{x}_{C+}$ where $s_C = s \cap U_C$.

We can write the CALWEIGHT estimator (4.1) as

$$\hat{y}_{d+} = y_{d\oplus\pi} + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T \hat{\mathbf{R}} \tag{4.2}$$

where $\hat{\mathbf{R}} = (\Sigma_s a_k \mathbf{z}_k \mathbf{x}_{Ck}^T)^{-1} (\Sigma_s a_k \mathbf{z}_k y_{dk})$. We note some of its properties:

i) In (4.2), we have $\mathrm{E}(y_{d\oplus\pi}) = y_{d+}$ and the expectation of the other term tends to zero, making $\hat{y}_{d+}$ design-consistent and very nearly design unbiased for y_{d+}.
ii) In practice, the same weight system, $w_k = a_k g_k$ for $k \in s$, is often used to produce estimates for any domain $U_d \subseteq U_C$; it is called a *uni-weight system* in Estevao and Särndal (1999).
iii) Estimator (4.1) is direct because the only y_k values used are for units inside the domain.

iv) Different choices of $\mathbf{z}_k$ give different weights g_k. The natural choice for $\mathbf{z}_k$, although not necessarily the best one, is to take $\mathbf{z}_k = \mathbf{x}_{Ck}$.

5. The Regression Fit Approach

The regression fit (REGFIT) approach starts by computing a sample-based regression vector, denoted $\hat{\mathbf{B}}$, for the regression of y on the auxiliary vector $\mathbf{x}$. The regression fit can be carried out at different levels, leading to different $\hat{\mathbf{B}}$. The estimator of y_{d+} is built by the principle

$$\hat{y}_{d+} = y_{d\oplus\pi} + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T\hat{\mathbf{B}} \tag{5.1}$$

This leads to a reduction in variance, compared to the simple HT estimator $y_{d\oplus\pi} = \Sigma_s a_k y_{dk}$, if there exists a negative correlation between the HT term $y_{d\oplus\pi}$ and the regression adjustment term $(\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T\hat{\mathbf{B}}$, which is a very nearly unbiased estimate of zero. The size of this reduction depends on (a) the given C-level, (b) the level of the regression fit that produces $\hat{\mathbf{B}}$, and (c) the correlation between y and $\mathbf{x}$. The C-level is fixed by the survey conditions and cannot be altered. It is better if the C-level is close to the d-level. Ideally, $U_C = U_d$ so that $\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi} = \mathbf{x}_{d+} - \mathbf{x}_{d\oplus\pi}$, assuming that $\mathbf{x}_{d+}$ is known. If the C-level is considerably above the d-level, the effect of the adjustment term may be small. Occasionally, it can lead to a variance even larger than that of $y_{d\oplus\pi}$, which uses no regression adjustment at all. The level at which the fit is carried out also has an impact on the variance of the estimator. We now consider different options for this regression level.

5.1. Regression fit at the domain level (REGFIT/DOM)

The motivation for this fit is that it recognizes differences between domains. One can argue that the domains have their own special characteristics and this local variation should be reflected in the underlying model. Therefore, we define a general regression fit at the domain level through the coefficient

$$\hat{\mathbf{B}}_{s_d} = \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}_{dk}^T\right)^{-1}\left(\sum\nolimits_s a_k \mathbf{z}_k y_{dk}\right) \tag{5.2}$$

where $y_{dk} = \delta_{dk} y_k$, $\mathbf{x}_{dk} = \delta_{dk}\mathbf{x}_k$ and $\mathbf{z}_k$ is an instrument vector of the same dimension as $\mathbf{x}_k$. The natural choice is $\mathbf{z}_k = \mathbf{x}_k$, leading to an ordinary least squares fit. The choice $\mathbf{z}_k = \mathbf{x}_k/c_k$, for specified positive constants c_k, corresponds to a generalized least squares fit. Other possibilities exist for $\mathbf{z}_k$. Setting $\hat{\mathbf{B}} = \hat{\mathbf{B}}_{s_d}$ in (5.1), we get the estimator

$$\hat{y}_{d+} = y_{d\oplus\pi} + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T\hat{\mathbf{B}}_{s_d} \tag{5.3}$$

We can express (5.3) as the weighted sum

$$\hat{y}_{d+} = \sum\nolimits_s a_k\{1 + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T\left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}_{dk}^T\right)^{-1}\mathbf{z}_k\} y_{dk}$$

Units outside the domain do not contribute to the sum, so (5.3) is a direct estimator.

This estimator is identical to the CALWEIGHT estimator (4.1) when the domain itself is the calibration group ($U_d = U_C$).

5.2. Regression fit at the full sample level (REGFIT/SAMPLE)

The motivation for this fit is to borrow strength by relying also on y-data from outside the domain itself to strengthen a potentially weak regression fit. To exploit this argument to its maximum extent, we should fit the regression of y on $\mathbf{x}$ at the full sample level, using the data $(\mathbf{x}_k, y_k)$ for $k \in s$. This gives

$$\hat{\mathbf{B}}_s = \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}_k^T\right)^{-1} \left(\sum\nolimits_s a_k \mathbf{z}_k y_k\right)$$

The natural choice for $\mathbf{z}_k$ is $\mathbf{z}_k = \mathbf{x}_k$, but other possibilities exist. Estimator (5.1) now becomes

$$\hat{y}_{d+} = y_{d\oplus\pi} + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T \hat{\mathbf{B}}_s \tag{5.4}$$

When expressed as a linearly weighted sum of y_k, (5.4) becomes

$$\hat{y}_{d+} = \sum\nolimits_s a_k \{\delta_{dk} + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}_k^T\right)^{-1} \mathbf{z}_k\} y_k$$

In general, this produces weights for all units $k \in s$, those inside as well as those outside the domain, making estimator (5.4) an indirect estimator that attempts to borrow strength by using y-data for the entire sample.

Other options exist. We can fit the regression at some intermediate level, following a pooling of data considered to come from similar domains. Such borrowing strength is considered in papers on small area estimation. An example occurs if the fit is carried out at the C-level. The estimator (5.1) then becomes

$$\hat{y}_{d+} = y_{d\oplus\pi} + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T \hat{\mathbf{B}}_{s_C}$$

where $\hat{\mathbf{B}}_{s_C}$ is given by (5.2) if we replace $y_{dk} = \delta_{dk} y_k$ and $\mathbf{x}_{dk} = \delta_{dk} \mathbf{x}_k$ by $y_{Ck} = \delta_{Ck} y_k$ and $\mathbf{x}_{Ck} = \delta_{Ck} \mathbf{x}_k$, respectively.

6. A General Class of Design-Based Domain Estimators

The CALWEIGHT and REGFIT approaches are built on different arguments. Both are sound in that they yield design-consistent and very nearly design unbiased estimators of y_{d+}. What is less evident is that they can differ considerably with respect to variance. We show this both by theoretical results (derivation of variances in Sections 6 to 9) and by empirical results (Monte Carlo simulation in Section 11).

Consider an auxiliary vector $\mathbf{x}_k$ for which we have the information (i) and (ii) in Section 3. We form a class of very nearly design unbiased estimators of y_{d+} that includes the CALWEIGHT and the REGFIT approaches:

$$\hat{y}_{d+} = y_{d\oplus\pi} + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T \hat{\mathbf{Q}}_{MLz} \tag{6.1}$$

with

$$\hat{\mathbf{Q}}_{MLz} = \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}_{Mk}^T\right)^{-1} \left(\sum\nolimits_s a_k \mathbf{z}_k y_{Lk}\right) \tag{6.2}$$

where $\mathbf{x}_{Mk} = \delta_{Mk}\mathbf{x}_k$, $y_{Lk} = \delta_{Lk}y_k$, and δ_{Mk} and δ_{Lk} are the respective identifiers associated with two new subpopulations, $U_M \subseteq U$ and $U_L \subseteq U$. By the principle used earlier, the identifier δ_{Mk} is 1 for all k inside U_M and 0 for all k outside, with a similar definition for δ_{Lk}. The fixed entities in the class (6.1) are the population U, the C-group U_C, and the domain U_d. Three factors that enter into $\hat{\mathbf{Q}}_{MLz}$ remain to be specified. They are the population levels U_M and U_L and the instrument $\mathbf{z}_k$.

It is easy to see that the class of estimators defined by (6.1) covers the CALWEIGHT estimator in Section 4 and the REGFIT estimators in Section 5. The CALWEIGHT estimator (4.1) is characterized by $U_M = U_C$ and $U_L = U_d$. The REGFIT estimator (5.1) is characterized by $U_L = U_M$. The REGFIT/DOM estimator (5.3) is obtained by $U_L = U_M = U_d$. The REGFIT/SAMPLE estimator (5.4) corresponds to $U_L = U_M = U$. REGFIT at the level of the fixed C-group is obtained by $U_L = U_M = U_C$.

The estimator (6.1) for y_{d+} contains the unbiased HT estimator $y_{d\oplus\pi}$ as one term. Why not use more of the available y-data in that term? Suppose that instead of $y_{d\oplus\pi}$ in (6.1) we use $(N_d/N_C)y_{C\oplus\pi}$, where $y_{C\oplus\pi} = \Sigma_s a_k y_{Ck}$ is the HT estimator of $y_{C+} = \Sigma_U y_{Ck}$, and N_d and N_C are the sizes of U_d and U_C, which we assume known. The resulting estimator is not design-consistent and thus beyond the scope of this article. It is biased for y_{d+} except in the unlikely circumstance that $y_{C+}/N_C = y_{d+}/N_d$.

The adjustment term $(\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T\hat{\mathbf{Q}}_{MLz}$ in (6.1) is a very nearly unbiased estimator of zero formed with the available auxiliary total $\mathbf{x}_{C+}$. Consequently, estimator (6.1) is design-consistent and very nearly design unbiased. We can measure the design-based variance of $\hat{y}_{d+}$, as discussed in the next section.

7. Design Measurability

The analysis steps for any design-based estimator include: (a) obtaining an (approximate) expression for its design-based variance, and, (b) deriving a design-consistent estimator of that variance from the sample data. Having carried out steps (a) and (b), we can make inferences about the finite population entirely on the basis of the randomization induced by the sampling design. The sampling literature then calls the procedure *design measurable*. Design measurability, a cornerstone of design-based reasoning, is possible when the bias ratio of the estimator tends to zero with increasing sample size. In the cases considered here, the bias ratio is $O(n^{-1/2})$. This has important implications for confidence intervals. In repeated samples, the interval centered on the point estimate and extending ± 1.96 times the estimated standard deviation has a coverage rate close to the nominal 95%, even for modest sample sizes. This has been borne out by many empirical studies.

Design measurability enables us to obtain, from the sample itself, an objective measure of the precision of the estimates. In the words of Hansen, Hurwitz, and Madow (1953, p 8–9), "the only insurance we have of the adequacy of the sample is the careful use of probability sampling methods and it requires probabilities of selection that are known." Cochran (1977, p 12–15, 160–162, 165–167), discusses design measurability

and the importance of a small bias ratio. He notes that a bias ratio that is $O(n^{-1/2})$ will not significantly perturb the coverage properties of design-based confidence intervals. These ideas are also important in Kish (1965), Hansen, Madow, and Tepping (1983) and Särndal, Swensson, and Wretman (1992).

We now establish design measurability for $\hat{y}_{d+}$ defined by (6.1). The error (the deviation from the target parameter) of $\hat{y}_{d+}$ is

$$\hat{y}_{d+} - y_{d+} = y_{d\oplus\pi} - y_{d+} + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T \hat{\mathbf{Q}}_{MLz} \tag{7.1}$$

where $\hat{\mathbf{Q}}_{MLz}$ is given by (6.2). An obstacle in the analysis of (7.1) is that $(\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T \hat{\mathbf{Q}}_{MLz}$ is a complex nonlinear term. We therefore find a close linear approximation to $\hat{y}_{d+} - y_{d+}$ and use it to derive the approximate variance of $\hat{y}_{d+}$. The nonlinearity of $(\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T \hat{\mathbf{Q}}_{MLz}$ ceases to be an obstacle if we can replace, with little error, the random $\hat{\mathbf{Q}}_{MLz}$ by a constant vector. This is done by centering $\hat{\mathbf{Q}}_{MLz}$ on the constant, nonrandom vector $\mathbf{Q}_{MLz} = (\Sigma_U \mathbf{z}_k \mathbf{x}_{Mk}^T)^{-1}(\Sigma_U \mathbf{z}_k y_{Lk})$ to which $\hat{\mathbf{Q}}_{MLz}$ converges in probability when the sample and the population increase in size. Now in (7.1) replace $\hat{\mathbf{Q}}_{MLz}$ by $\mathbf{Q}_{MLz} + (\hat{\mathbf{Q}}_{MLz} - \mathbf{Q}_{MLz})$ and rearrange terms. We get

$$\hat{y}_{d+} - y_{d+} = e_{C\oplus\pi} - e_{C+} - (\mathbf{x}_{C\oplus\pi} - \mathbf{x}_{C+})^T (\hat{\mathbf{Q}}_{MLz} - \mathbf{Q}_{MLz}) \tag{7.2}$$

where $e_{C\oplus\pi} = \Sigma_s a_k e_{Ck}$ and $e_{C+} = \Sigma_U e_{Ck}$ with

$$e_{Ck} = y_{dk} - \mathbf{x}_{Ck}^T \mathbf{Q}_{MLz} \tag{7.3}$$

It follows that $e_{Ck} = y_k - \mathbf{x}_k^T \mathbf{Q}_{MLz}$ for $k \in U_d$; $e_{Ck} = -\mathbf{x}_k^T \mathbf{Q}_{MLz}$ for $k \in U_C - U_d$; and $e_{Ck} = 0$ for all $k \notin U_C$. It is clearly understood that e_{Ck} is also a function of the domain, but for simplicity of notation we only include the subscript for the C-group since the domain U_d is given and we examine the properties of $\hat{y}_{d+}$ for different C-groups. We progress from (7.1) to (7.2) by centering $\hat{\mathbf{Q}}_{MLz}$ on its constant counterpart $\mathbf{Q}_{MLz}$. This creates a term of lower order of importance: In (7.2), the two differences $e_{C\oplus\pi} - e_{C+}$ and $\mathbf{x}_{C\oplus\pi} - \mathbf{x}_{C+}$ have (i) a zero expectation, and (ii) the same order in probability, because when multiplied by N^{-1}, each is $O_p(n^{-1/2})$ under general conditions. The term $(\hat{\mathbf{Q}}_{MLz} - \mathbf{Q}_{MLz})$ is close to $\mathbf{0}$ to the same order. Then the product $N^{-1}(\mathbf{x}_{C\oplus\pi} - \mathbf{x}_{C+})^T(\hat{\mathbf{Q}}_{MLz} - \mathbf{Q}_{MLz})$ is $O_p(n^{-1})$, thus of lower order than (and usually negligible compared to) $N^{-1}(e_{C\oplus\pi} - e_{C+})$, and the latter term alone provides the desired close linear approximation:

$$N^{-1}(\hat{y}_{d+} - y_{d+}) = N^{-1}(e_{C\oplus\pi} - e_{C+}) + O_p(n^{-1}) \approx N^{-1}(e_{C\oplus\pi} - e_{C+})$$

As a result, the bias and the variance of $\hat{y}_{d+}$ can now be closely approximated by the easily derived counterparts for the linear statistic $e_{C\oplus\pi}$. Because $E(e_{C\oplus\pi}) = e_{C+}$, the bias of $\hat{y}_{d+}$ is approximately zero. An exact expression for the bias of $\hat{y}_{d+}$ is, from (7.2),

$$\text{Bias}(\hat{y}_{d+}) = E(\hat{y}_{d+}) - y_{d+} = -E\{(\mathbf{x}_{C\oplus\pi} - \mathbf{x}_{C+})^T(\hat{\mathbf{Q}}_{MLz} - \mathbf{Q}_{MLz})\}$$

For the bias we have $N^{-1}\text{Bias}(\hat{y}_{d+}) = O(n^{-1})$, and for the variance $N^{-2}\text{Var}(\hat{y}_{d+}) = O(n^{-1})$. Thus the bias ratio of $\hat{y}_{d+}$ is $O(n^{-1/2})$, and an essential requirement of design-based inference is thereby met. Even for modest sample sizes n, this small bias does not seriously perturb the validity of a design-based confidence interval. The interpretation of these asymptotics is as follows. There is a series of growing populations U and growing

samples s. The subpopulations U_d, U_C, U_M and U_L, also grow, at constant rates. The main conclusions of this section are summarized in the following result.

Result 7.1 The estimator $\hat{y}_{d+}$ given by (6.1) is design measurable. Its bias ratio is $O(n^{-1/2})$ under general conditions. Its design-based variance, $\mathrm{Var}(\hat{y}_{d+})$, is closely approximated by the asymptotic variance

$$\mathrm{Var}(e_{C\oplus\pi}) = \sum\sum_U \left(\frac{a_k a_l}{a_{kl}} - 1\right) e_{Ck} e_{Cl}$$

where e_{Ck} is given by (7.3), $a_{kl} = 1/\pi_{kl}$, π_{kl} is the joint inclusion probability of units k and l under the given design, $a_{kk} = 1/\pi_k$ and $\Sigma\Sigma_U$ denotes the double sum $\Sigma_{k\in U}\Sigma_{l\in U}$.

Our simulations reported in Section 11 indicate that for domains and samples of rather modest sizes, the bias is small and that the approximation in Result 7.1 succeeds well in measuring the variance. The estimator $\hat{y}_{d+}$ is design-consistent and very nearly design unbiased, in that its bias ratio tends to zero as strongly as $O(n^{-1/2})$.

8. Achieving Minimum Asymptotic Variance

In constructing $\hat{y}_{d+}$ defined by (6.1) we used information about a fixed C-group, U_C, with its known auxiliary vector total $\mathbf{x}_{C+}$. Now (6.1) depends, through $\hat{\mathbf{Q}}_{MLz}$ given by (6.2), on the two levels U_M and U_L and on the instrument vector $\mathbf{z}_k$. Section 7 showed that $\mathrm{Var}(\hat{y}_{d+}) \approx \mathrm{Var}(e_{C\oplus\pi})$, where $e_{C\oplus\pi} = \Sigma_s a_k e_{Ck}$ with $e_{Ck} = y_{dk} - \mathbf{x}_{Ck}^T \mathbf{Q}_{MLz}$. We now find the vector $\mathbf{Q}_{MLz}$ that minimizes $\mathrm{Var}(e_{C\oplus\pi})$. That is, we look for the optimal choices of U_M, U_L and $\mathbf{z}_k$. This search is facilitated by noting the presence here of two transformed variables, the vector variable $\mathbf{x}_C$ with value $\mathbf{x}_{Ck} = \delta_{Ck}\mathbf{x}_k$ and the domain variable y_d with value $y_{dk} = \delta_{dk} y_k$, both defined for every unit $k \in U$. In (6.1), $\mathbf{x}_{C\oplus\pi}$ and $y_{d\oplus\pi}$ are the unbiased HT estimators of the totals of $\mathbf{x}_{Ck}$ and y_{dk} respectively. After dropping the lower order term in (7.2), we have a linear statistic that approximates $\hat{y}_{d+}$, namely,

$$\hat{y}_{d+}^0 = e_{d\oplus\pi} + \mathbf{x}_{C+}^T \mathbf{Q}_{MLz} = y_{d\oplus\pi} - (\mathbf{x}_{C\oplus\pi} - \mathbf{x}_{C+})^T \mathbf{Q}_{MLz}$$

The minimization of the variance of $\hat{y}_{d+}^0$ proceeds as in Montanari (1987), although our variables are different. Because $\mathbf{Q}_{MLz}$ is a constant vector, and $y_{d\oplus\pi}$ and $\mathbf{x}_{C\oplus\pi}$ are HT estimators, we get

$$\mathrm{Var}(\hat{y}_{d+}^0) = \mathrm{Var}(y_{d\oplus\pi}) + \mathbf{Q}_{MLz}^T \mathrm{Var}(\mathbf{x}_{C\oplus\pi}) \mathbf{Q}_{MLz} - 2\mathrm{Cov}(\mathbf{x}_{C\oplus\pi}, y_{d\oplus\pi})^T \mathbf{Q}_{MLz} \tag{8.1}$$

where $\mathrm{Var}(\mathbf{x}_{C\oplus\pi}) = \Sigma\Sigma_U(\frac{a_k a_l}{a_{kl}} - 1)\mathbf{x}_{Ck}\mathbf{x}_{Cl}^T$ and $\mathrm{Cov}(\mathbf{x}_{C\oplus\pi}, y_{d\oplus\pi}) = \Sigma\Sigma_U(\frac{a_k a_l}{a_{kl}} - 1)\mathbf{x}_{Ck} y_{dl}$. The minimum of the quadratic form (8.1) with respect to $\mathbf{Q}_{MLz}$ is realized for $\mathbf{Q}_{MLz} = \mathbf{Q}_{Cdz}$ where

$$\mathbf{Q}_{Cdz} = \{\mathrm{Var}(\mathbf{x}_{C\oplus\pi})\}^{-1} \mathrm{Cov}(\mathbf{x}_{C\oplus\pi}, y_{d\oplus\pi}) \tag{8.2}$$

assuming $\mathrm{Var}(\mathbf{x}_{C\oplus\pi})$ is nonsingular. Thus, optimal choices are $U_M = U_C$, $U_L = U_d$ and $\mathbf{z}_k = \mathbf{z}_{Uk}$ where

$$\mathbf{z}_{Uk} = \sum_{l\in U} \left(\frac{a_k a_l}{a_{kl}} - 1\right) \mathbf{x}_{Cl} \tag{8.3}$$

Because $\mathbf{Q}_{Cdz}$ is unknown, $\hat{y}^0_{d+}$ with $\mathbf{Q}_{MLz} = \mathbf{Q}_{Cdz}$ is not a proper estimator. In practice, we must replace it by a sample-based vector. Let $\hat{\mathbf{Q}}_{Cdz}$ be the sample-based analogue of $\mathbf{Q}_{Cdz}$, found by replacing the variance and the covariance on the right-hand side of (8.2) by their usual sample-based (and unbiased) counterparts. Consequently, $\hat{\mathbf{Q}}_{Cdz} = (\Sigma_s a_k \mathbf{z}_k \mathbf{x}_{Ck}^T)^{-1} \times (\Sigma_s a_k \mathbf{z}_k y_{dk})$ with $\mathbf{z}_k = \mathbf{z}_{sk}$ given by

$$\mathbf{z}_{sk} = a_k^{-1} \sum\nolimits_{l \in s} (a_k a_l - a_{kl}) \mathbf{x}_{Cl} \tag{8.4}$$

This leads us to the following result.

Result 8.1 For the given auxiliary information (i) and (ii) in Section 3, the asymptotically optimal estimator of y_{d+} in the class (6.1) is

$$\hat{y}_{d+} = y_{d\oplus\pi} + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T \hat{\mathbf{Q}}_{Cdz}$$

where $\hat{\mathbf{Q}}_{Cdz} = (\Sigma_s a_k \mathbf{z}_k \mathbf{x}_{Ck}^T)^{-1} (\Sigma_s a_k \mathbf{z}_k y_{dk})$ with $\mathbf{z}_k = \mathbf{z}_{sk} = a_k^{-1} \Sigma_{l \in s} (a_k a_l - a_{kl})\, \mathbf{x}_{Cl}$.

This is a CALWEIGHT estimator of the form given by (4.2). None of the REGFIT options in the class (6.1) has a smaller asymptotic variance unless $U_C = U_d$, in which case CALWEIGHT and REGFIT/DOM produce the same asymptotically optimal estimator for $\mathbf{z}_{sk}$ given by (8.4).

This raises the question whether software such as GES (Statistics Canada) and CLAN97 (Statistics Sweden), rely on the optimal procedure in Result 8.1. The answer is "they come fairly close." They compute estimator (6.1) with $U_M = U_C$, $U_L = U_d$ and $\mathbf{z}_k = \mathbf{x}_{Ck}$. Of these, $U_M = U_C$ and $U_L = U_d$ are optimal choices, but $\mathbf{z}_k = \mathbf{x}_{Ck}$ is not. The loss of efficiency may be small in most cases, but exceptions could exist.

In estimation for the whole population U, the asymptotically optimal estimator has been carefully examined in, for example, Casady and Valliant (1993), Montanari (1998, 2000), and Montanari and Ranalli (2002). It is known to be unstable, especially for designs more complex than SRS. Here we encounter the asymptotically optimal estimator in the context of domain estimation. The discussion in the cited references is relevant here too. The choice $\mathbf{z}_k = \mathbf{z}_{sk} = a_k^{-1} \sum_{l \in s} (a_k a_l - a_{kl})\, \mathbf{x}_{Cl}$ can lead to an unstable estimator. A prudent approach is to use $\mathbf{z}_k = \mathbf{x}_{Ck}$ in all cases.

9. An Analysis of Stratified Simple Random Sampling

Expressions (8.2) and (8.3) for the optimal $\mathbf{Q}_{Cdz}$ and $\mathbf{z}_k$ depend on the sampling design. The cumbersome double sums in (8.2) simplify for some designs of practical interest. These include Poisson sampling and Stratified Simple Random Sampling (STSRS). For Poisson sampling, $\mathbf{Q}_{Cdz}$ simplifies because $a_{kl} = a_k a_l$ for all $k \neq l$, so only off-diagonal terms remain in the double sums in (8.2). Almost as simple is STSRS. Because of its importance in practice, we illustrate the optimal form of $\mathbf{Q}_{Cdz}$ and $\mathbf{z}_k$ for STSRS.

Suppose that the population U of size N is divided into H strata, $U_h, h = 1, \ldots, H$. For U_h, let the sampling rate be $f_h = n_h/N_h$, $N = \Sigma_{h=1}^H N_h$, and set $K_h = \frac{N_h}{N_h - 1}(\frac{1}{f_h} - 1) \approx \frac{1}{f_h} - 1$.

Using (8.2) and (8.3), we obtain after some algebra

$$\mathbf{Q}_{Cdz} = \left(\sum_{h=1}^{H} K_h \left\{\sum_{U_h} (\mathbf{x}_{Ck} - \bar{\mathbf{x}}_{CU_h})\mathbf{x}_{Ck}^T\right\}\right)^{-1} \left(\sum_{h=1}^{H} K_h \left\{\sum_{U_h} (\mathbf{x}_{Ck} - \bar{\mathbf{x}}_{CU_h}) y_{dk}\right\}\right) \tag{9.1}$$

The Q-vector that minimizes the asymptotic variance under STSRS is therefore characterized by $U_M = U_C$, $U_L = U_d$ and $\mathbf{z}_k = \mathbf{z}_{Uk} = K_h(\mathbf{x}_{Ck} - \bar{\mathbf{x}}_{CU_h})$ for $k \in U_h$, where $\bar{\mathbf{x}}_{CU_h} = \Sigma_{U_h}\mathbf{x}_{Ck}/N_h$ is the mean of $\mathbf{x}_{Ck} = \delta_{Ck}\mathbf{x}_k$ in stratum U_h. Unless U_C is identical to U_h, $\bar{\mathbf{x}}_{CU_h}$ differs from $\bar{\mathbf{x}}_{U_{Ch}} = \Sigma_{U_{Ch}}\mathbf{x}_k/N_{Ch}$, the mean of $\mathbf{x}_k$ for the N_{Ch} units in $U_{Ch} = U_C \cap U_h$. These two means are related by $\bar{\mathbf{x}}_{CU_h} = P_{Ch}\bar{\mathbf{x}}_{U_{Ch}}$, where $P_{Ch} = N_{Ch}/N_h$. The asymptotically optimal estimator of the domain total is

$$\hat{y}_{d+} = y_{d\oplus\pi} + (\mathbf{x}_{C+} - \mathbf{x}_{C\oplus\pi})^T \hat{\mathbf{Q}}_{Cdz} \tag{9.2}$$

It is obtained by replacing $\mathbf{Q}_{Cdz}$ given by (9.1) by its sample-based analogue,

$$\hat{\mathbf{Q}}_{Cdz} = \left(\sum_{h=1}^{H} \breve{K}_h \left\{\sum_{s_h} (\mathbf{x}_{Ck} - \bar{\mathbf{x}}_{Cs_h})\mathbf{x}_{Ck}^T\right\}\right)^{-1} \left(\sum_{h=1}^{H} \breve{K}_h \left\{\sum_{s_h} (\mathbf{x}_{Ck} - \bar{\mathbf{x}}_{Cs_h}) y_{dk}\right\}\right) \tag{9.3}$$

where $\breve{K}_h = \frac{n_h}{n_h - 1}\frac{1}{f_h}(\frac{1}{f_h} - 1) \approx \frac{1}{f_h}(\frac{1}{f_h} - 1)$ and $\bar{\mathbf{x}}_{Cs_h}$ is the mean of $\mathbf{x}_{Ck} = \delta_{Ck}\mathbf{x}_k$ in the simple random sample s_h from U_h. Properties of (9.2) are: (i) it is a direct estimator, and (ii) it has the form of the CALWEIGHT estimator (4.2) with the instrument vector $\mathbf{z}_k = \mathbf{z}_{sk}$ obtained from (8.4) as

$$\mathbf{z}_{sk} = \frac{n_h}{n_h - 1}\left(\frac{1}{f_h} - 1\right)(\mathbf{x}_{Ck} - \bar{\mathbf{x}}_{Cs_h}) \tag{9.4}$$

for $k \in s_h$, $h = 1, \ldots, H$. The results of this section are used for the simulation in Section 11 in the special case $H = 1$. Formulas (9.1) and (9.3) have familiar appearances. Rao (1994) and Montanari (1998, 2000) show them in the more usual form when $\mathbf{x}_k$ and y_k take the place of our variables $\mathbf{x}_{Ck}$ and y_{dk}.

10. Domains That Intersect Several Calibration Groups

The results in Sections 4 to 9 concern a domain of interest wholly contained in one C-group. Often in practice, a domain of interest cuts across several C-groups, each having a known auxiliary vector total. The results continue to apply if the following modifications are made.

Let the population U be composed of I C-groups, denoted U_{C_i}, $i = 1, \ldots, I$. The domain of interest U_d may intersect several of them. Let $\delta_{C_i k} = 1$ if $k \in U_{C_i}$ and 0 otherwise, for $i = 1, \ldots, I$. For simplicity suppose the auxiliary value x_k is scalar, but more generally, x_k can be a vector. Define $x_{C_i k} = \delta_{C_i k} x_k$.

The auxiliary information is stated as follows, for $i = 1, \ldots, I$:

i) For U_{C_i} the auxiliary total $x_{C_i+} = \Sigma_{U_{C_i}} x_k = \Sigma_U x_{C_i k}$ is known
ii) For every $k \in s$, the value x_k and membership or not of k in U_{C_i} are known.

The theory in Sections 4 to 9 applies if we take $U_C = U$ and $\mathbf{x}_{Ck} = \mathbf{x}_k$, where $\mathbf{x}_k = (x_{C_1k}, \ldots, x_{C_ik}, \ldots, x_{C_Ik})^T$, of dimension I. Note that $\mathbf{x}_+ = \Sigma_U \mathbf{x}_k = (x_{C_1+}, \ldots, x_{C_i+}, \ldots, x_{C_I+})^T$, the vector of known x-totals.

To illustrate, suppose that STSRS is used (with notation as in Section 9) in such a way that each stratum is identical to a C-group, a case often found in practice. That is, each stratum U_h coincides exactly with one of the calibration groups U_{C_i}. We have $I = H$, and $\mathbf{x}_k = (\delta_{1k}x_k, \ldots, \delta_{hk}x_k, \ldots, \delta_{Hk}x_k)^T$, where δ_{hk} is now the identifier of stratum U_h. The required information (i) consists of the stratum totals $\Sigma_{U_h}x_k$, for $h = 1, \ldots, H$. A derivation using (9.3) shows that the weights in the minimum variance estimator (9.2) are $w_k = a_k g_k$ with $a_k = N_h/n_h = 1/f_h$ and $g_k = 1 + D_h(x_k - \bar{x}_{s_h})$ for $k \in s_h$, with $D_h = (\bar{x}_{U_h} - \bar{x}_{s_h})/S^2_{xs_h}$, $S^2_{xs_h} = \Sigma_{s_h}(x_k - \bar{x}_{s_h})^2/n_h$, $\bar{x}_{s_h} = \Sigma_{s_h}x_k/n_h$ and $\bar{x}_{U_h} = \Sigma_{U_h}x_k/N_h$. The asymptotically optimal estimator of y_{d+} becomes

$$\hat{y}_{d+} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum\nolimits_{s_h} y_{dk} + \sum_{h=1}^{H} N_h(\bar{x}_{U_h} - \bar{x}_{s_h})\hat{R}_h \tag{10.1}$$

with $\hat{R}_h = (1/n_h)\Sigma_{s_h}(x_k - \bar{x}_{s_h})y_{dk}/S^2_{xs_h}$. For the same information, no other estimator in our class can have a smaller asymptotic variance. Fitting the regression with the aid of the whole y-data set is tempting, at first sight. But for estimating y_{d+} this gives the less efficient (although still design-consistent) alternative

$$\hat{y}_{d+} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum\nolimits_{s_h} y_{dk} + \sum_{h=1}^{H} N_h(\bar{x}_{U_h} - \bar{x}_{s_h})\hat{B}_h \tag{10.2}$$

with $\hat{B}_h = (1/n_h)\Sigma_{s_h}(x_k - \bar{x}_{s_h})(y_k - \bar{y}_{s_h})/S^2_{xs_h}$. One notes that synthetic estimators such as

$$\hat{y}_{d+} = \sum_{h=1}^{H} x_{dh+}\hat{B}_h \quad \text{or} \quad \hat{y}_{d+} = \sum_{h=1}^{H} N_{dh}\{\bar{y}_{s_h} + (\bar{x}_{U_h} - \bar{x}_{s_h})\hat{B}_h\}$$

do not qualify under the auxiliary information requirements (i) and (ii). They require the more extensive auxiliary information $x_{dh+} = \Sigma_{U_{dh}}x_k$, $h = 1, \ldots, H$ (for the former), and N_{dh} and $x_{h+} = \Sigma_{U_h}x_k$, $h = 1, \ldots, H$ (for the latter), where $U_{dh} = U_d \cap U_h$ and N_{dh} its size. Their variance may be lower than that of (10.1) or (10.2) but their mean squared error may be larger because of the squared bias component.

11. Simulation

This section describes our simulation for a given domain $U_d \subset U$. We study estimators $\hat{y}_{d+}$ in the class defined by (6.1). This class is characterized by the fixed set of auxiliary information given by (i) and (ii) in Section 3. For this fixed information, the members of the class correspond to the different choices of U_M, U_L and $\mathbf{z}_k$. The CALWEIGHT estimators, obtained when $U_M = U_C$ and $U_L = U_d$, have the form (4.2). The optimal CALWEIGHT estimator is the one with $\mathbf{z}_k = \mathbf{z}_{sk}$ specified by (8.4). The various REGFIT alternatives in Section 5 are obtained when $U_M = U_L$. When $U_C = U_d$, the CALWEIGHT and REGFIT/DOM estimators are the same.

One objective of the simulation is to ascertain whether the CALWEIGHT estimator has smaller variance than all REGFIT alternatives when $\mathbf{z}_k = \mathbf{z}_{sk}$ and $U_C \supset U_d$. The backing for this supposition is the asymptotic theory in Section 8, but we expect to find it to hold

also for the finite conditions in this simulation. Other questions that can be at least partially answered by simulation (in contrast to analytic derivations) are:

Q1. Is the variance of the optimal CALWEIGHT estimator much smaller than that of the REGFIT alternatives, or are differences only negligible?
Q2. To what extent is the variance of the optimal CALWEIGHT estimator sensitive to the level of the C-group U_C that contains the given domain U_d? As Estevao and Särndal (1999) note, the asymptotic variance is minimal when $U_C = U_d$ and it increases steadily as U_C expands from U_d to U.

Our conclusions in regard to Q1 and Q2 are only indicative. More complete answers about these issues would require simulations on many different populations, which is beyond the scope of this article.

Our simulation consisted in drawing repeated SRS samples of size $n = 1,500$ from an artificially generated population of size $N = 5,000$. We used the following two-step procedure to create a finite population consisting of $N = 5,000$ pairs (x_k, y_k), $k = 1, 2, \ldots, 5,000$, where Gamma(a,b) refers to the gamma distributed random variable with density function $f(x) = \{\Gamma(a)b^a\}^{-1}x^{a-1}e^{-x/b}$ for $x > 0$:

(1) First, create the 5,000 x_k values as independent realizations of Gamma(2,5). Consequently, the mean of the resulting x_k values will be roughly equal to the theoretical mean, $\mu_x = ab = 10$, and their variance roughly equal to $ab^2 = 50$.
(2) Then, given x_k, create a corresponding value y_k as one realization of Gamma(A_k, B_k), where the parameters A_k and B_k are chosen so that y_k conditionally on x_k has expected value $\mu_{y_k|x_k} = \alpha + \beta x_k + Kx_k(x_k - \mu_x)(x_k - 3\,\mu_x)$ and variance $\sigma^2_{y_k|x_k} = \sigma^2 x_k$; $k = 1, 2, \ldots, 5,000$. That is, $A_k = (\mu_{y_k|x_k})^2/\sigma^2_{y_k|x_k}$ and $B_k = \sigma^2_{y_k|x_k}/\mu_{y_k|x_k}$.

We used $\alpha = 20$, $\beta = 1$, $K = 0.001$, $\sigma^2 = 5$, and we have $\mu_x = 10$ by the first step. The mean and the standard deviation of the 5,000 x-values were 10.01 and 6.97, respectively. The corresponding moments for the 5,000 y-values were 30.01 and 9.80, respectively. The relationship between y and x is slightly curved as a result of using a K different from but near zero. This is to avoid an argument that some of the simulation results may happen just because of a population model with a perfect linear regression. Figure 1 shows the scatter plot of the 5,000 points and the least squares linear regression line with slope 0.95 and intercept 20.54. It clearly indicates the nonlinear pattern of the bivariate plot. The properties that we wish to illustrate are independent of the form of the relation between y and x. The correlation coefficient between y and x, computed on the 5,000 generated pairs (x_k, y_k), is 0.67. We carried out the same simulation on populations generated with other values of K close to 0. The principal conclusions are the same, so those simulation results are not reported here.

Given the population $U = U_{5000}$, we then proceeded to create four other C-groups, the smallest of which is the domain of study, U_d. We first created $U_C = U_{2500}$ as an SRS selection of size 2,500 from U_{5000}. Then, we obtained $U_C = U_{1000}$ as an SRS selection of size 1,000 from U_{2500}, $U_C = U_{600}$ as an SRS selection of size 600 from U_{1000}, and finally, $U_C = U_{500} = U_d$ as an SRS selection of size 500 from U_{600}. By this construction, the domain $U_d = U_{500}$ is entirely contained in each of the C-groups. Furthermore, U_{5000},

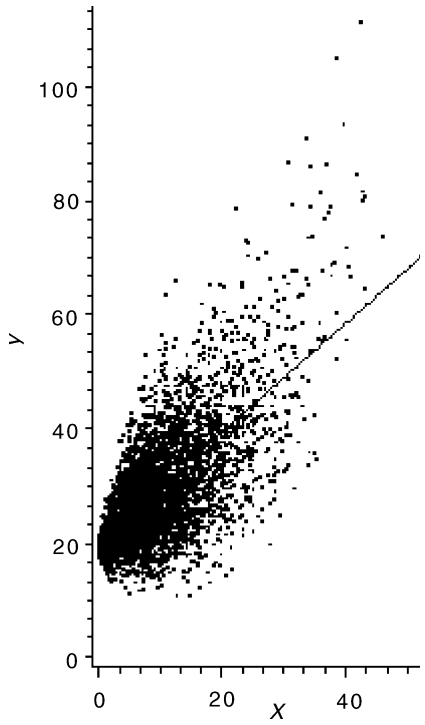

Fig. 1. Scatter plot and regression fit of the simulated population

U_{2500}, U_{1000}, U_{600}, and U_{500} have similar means and variances for both x and y. This setup provides ideal conditions for borrowing strength from the larger groups to the domain $U_d = U_{500}$, which is 10% of the population U_{5000}. The expected domain sample size in this simulation is 150. The reason that the chosen five U_C are "unevenly spaced" is that a considerable increase in the variance of $\hat{y}_{d+}$ is anticipated as soon as U_C starts to distance itself from the fixed U_d. To observe this, we included $U_C = U_{600}$ even though it is close to the domain $U_d = U_{500}$. The shape of the distribution of the y_k values is without consequence. The results were confirmed for other distributions.

The population $U = U_{5000}$ and the domain of interest $U_d = U_{500}$ are fixed throughout the simulation. The domain y-total $y_{d+} = \Sigma_{U_{500}} y_k$ is always the target of estimation. Estimator (6.1) depends on a number of factors. In our simulation, we considered the following factors and values, where $\mathbf{x}_{Ck} = \delta_{Ck}\mathbf{x}_k$ and $\bar{\mathbf{x}}_{Cs} = \Sigma_s \mathbf{x}_{Ck}/n$.

(1) the auxiliary vector $\mathbf{x}_k$: $\mathbf{x}_k = x_k$ and $\mathbf{x}_k = (1, x_k)^T$.
(2) the instrument vector $\mathbf{z}_k$: $\mathbf{z}_k = \mathbf{z}_{sk} = \mathbf{x}_{Ck} - \bar{\mathbf{x}}_{Cs}$, $\mathbf{z}_k = \mathbf{x}_{Ck}$ and $\mathbf{z}_k = \mathbf{x}_k$.
(3) the calibration group U_C : $U_C = U_{5000}$, $U_C = U_{2500}$, $U_C = U_{1000}$, $U_C = U_{600}$ and $U_C = U_{500}$.
(4) the subpopulation U_M : $U_M = U_{5000}$, $U_M = U_{2500}$, $U_M = U_{1000}$, $U_M = U_{600}$ and $U_M = U_{500}$.
(5) the subpopulation U_L : $U_L = U_{5000}$, $U_L = U_{2500}$, $U_L = U_{1000}$, $U_L = U_{600}$ and $U_L = U_{500}$.

The auxiliary vector $\mathbf{x}_k$ establishes an important classification of the estimators. Each of the two auxiliary vectors creates a family of estimators, and in each family we can determine an asymptotically optimal estimator for any given sampling design, such as the SRS design used here. For the family of estimators with $\mathbf{x}_k = x_k$ we need to know the C-group auxiliary total x_{C+}. The other family has $\mathbf{x}_k = (1, x_k)^T$ and we must know the size N_C of the group U_C, in addition to x_{C+}.

For a given $\mathbf{x}_k$, the optimal CALWEIGHT estimator for SRS is the one with the instrument vector $\mathbf{z}_k = \mathbf{z}_{sk} = \mathbf{x}_{Ck} - \bar{\mathbf{x}}_{Cs}$. This follows from (9.4) since $(n/(n-1)) \times (1/f - 1)$ is constant and cancels out in the case of $H = 1$ stratum. We also consider the simple "obvious" choices $\mathbf{z}_k = \mathbf{x}_{Ck}$ and $\mathbf{z}_k = \mathbf{x}_k$. Thus our simulation involves six sets of estimators $\hat{y}_{d+}$, produced by the two choices of $\mathbf{x}_k$ and the three choices of $\mathbf{z}_k$.

We use the same five choices for U_M, U_L and U_C. Consequently, for each fixed triple $(\mathbf{x}_k, \mathbf{z}_k, U_C)$, we have $5 \times 5 = 25$ well-defined estimators $\hat{y}_{d+}$, except when $\mathbf{x}_k = (1, x_k)^T$, $\mathbf{z}_k = \mathbf{z}_{sk}$ and $U_C = U = U_{5000}$. Then the 2×2 matrix $(\Sigma_s a_k \mathbf{z}_k \mathbf{x}_{Mk}^T)$ is singular, and $\hat{y}_{d+}$ is undefined, because the first component of $\mathbf{z}_k = \mathbf{z}_{sk} = \mathbf{x}_{Ck} - \bar{\mathbf{x}}_{Cs} = \mathbf{x}_k - \bar{\mathbf{x}}_s$ is zero for all k. Two facts have a bearing on this exceptional case:

i) For SRS, $\mathbf{x}_k = (1, x_k)^T$, $U_C \subset U$, $U_M \subseteq U_C$ and $U_L \subseteq U_C$, one can show that $\hat{y}_{d+}$ given by (6.1) is the same for all three instruments $\mathbf{z}_k = \mathbf{z}_{sk} = \mathbf{x}_{Ck} - \bar{\mathbf{x}}_{Cs}$, $\mathbf{z}_k = \mathbf{x}_{Ck}$ and $\mathbf{z}_k = \mathbf{x}_k$. This holds for any U_C that is a proper subset of U, but not when $U_C = U$. Also, the property does not hold for $\mathbf{x}_k = x_k$.

ii) In view of the singularity of the 2×2 matrix, we can remove the auxiliary "1" from $\mathbf{x}_k = (1, x_k)^T$, leaving $\mathbf{x}_k = x_k$. Then, for SRS and $U_C = U = U_M$, $\hat{y}_{d+}$ with $\mathbf{x}_k = x_k$ and the corresponding optimal $\mathbf{z}_k = z_{sk} = x_{Ck} - \bar{x}_{Cs}$ is identical to $\hat{y}_{d+}$ with $\mathbf{x}_k = (1, x_k)^T$ and $\mathbf{z}_k = \mathbf{x}_k$. This follows from Montanari (1998, 2000).

In view of (i) and (ii), it is natural in the exceptional case $\mathbf{x}_k = (1, x_k)^T$, $\mathbf{z}_k = \mathbf{z}_{sk} = \mathbf{x}_{Ck} - \bar{\mathbf{x}}_{Cs}$, $U_C = U_{5000}$ to "import" the simulation results for the case $\mathbf{x}_k = (1, x_k)^T$, $\mathbf{z}_k = \mathbf{x}_k$, $U_C = U_{5000}$. That is, we fill the gap caused by the undefined $\hat{y}_{d+}$ for the former case with the well-defined $\hat{y}_{d+}$ for the latter case.

In the simulation, we drew $M = 100,000$ independent SRS samples of size $n = 1,500$ from $U = U_{5000}$. For each of these samples, we computed every estimator within each set of 25. Let $\hat{y}_{d+,j}$ represent the value of $\hat{y}_{d+}$ (one of the estimators in the simulation) in sample j, for $j = 1, \ldots, M = 100,000$. Then for each estimator, we computed the following statistics:

- the Monte Carlo expectation, $\frac{1}{M}\Sigma_{j=1}^{M} \hat{y}_{d+,j}$
- the Monte Carlo variance (MCVar), $\frac{1}{M-1}\Sigma_{j=1}^{M} \left(\hat{y}_{d+,j} - \frac{1}{M}\Sigma_{j=1}^{M} \hat{y}_{d+,j}\right)^2$
- the asymptotic variance (AVar) of $\hat{y}_{d+}$

$$N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{\sum_U (e_{Ck} - \bar{e}_{CU})^2}{N-1} \quad \text{with} \tag{11.1}$$

$$e_{Ck} = y_{dk} - \mathbf{x}_{Ck}^T \mathbf{Q}_{MLz} \quad \text{and} \quad \bar{e}_{CU} = \frac{\sum_U e_{Ck}}{N}$$

Formula (11.1) is derived by applying Result 7.1 to the special case of SRS.

A measure of the bias of any of the estimators is obtained as the difference between the Monte Carlo expectation and the true value of the domain total. As expected by theory, the square of this difference was always negligible in comparison with the Monte Carlo variance, thus contributing insignificantly to the mean squared error. Consequently, our tables do not show figures for the bias.

For every one of the five groups U_C, Table 1 shows the MCVar of the 25 estimators $\hat{y}_{d+}$ in (6.1) for $\mathbf{x}_k = (1, x_k)^T$ and the optimal instrument $\mathbf{z}_k = \mathbf{z}_{sk} = \mathbf{x}_{Ck} - \bar{\mathbf{x}}_{Cs}$. For the same specifications, Table 2 shows the AVar of $\hat{y}_{d+}$, computed by (11.1) with $\mathbf{z}_k$ in $\mathbf{Q}_{MLz}$ given by $\mathbf{z}_k = \mathbf{z}_{Uk} = \mathbf{x}_{Ck} - \bar{\mathbf{x}}_{CU} = \mathbf{x}_{Ck} - (\Sigma_U x_{Ck}/N)$. Table 2 also shows the ratio MCVar/AVar. We cannot compute the estimator for the case $U_C = U = U_{5000}$ in Tables 1 and 2 because of the singularity of $(\Sigma_s a_k \mathbf{z}_{sk} \mathbf{x}_{Mk}^T)$. The matrix $(\Sigma_U \mathbf{z}_{Uk} \mathbf{x}_{Mk}^T)$ is also singular. However, the rationale given earlier allows us to justify using the results of $\mathbf{z}_k = \mathbf{x}_k$ for this exceptional case.

For each U_C, the cells of principal interest, out of the 25 in Tables 1 and 2, are those with $U_M = U_C$ and $U_L = U_d$ (the CALWEIGHT estimator), and the five cells on the diagonal having $U_M = U_L$ (the REGFIT estimators). Other cells are included for comparison only; these estimators belong in the family (6.1) but have no clear interpretation. Results for other combinations of $\mathbf{x}_k$ and $\mathbf{z}_k$ are not reported here to save space, but some comments are given below.

11.1. Comparing the Monte Carlo variance to the asymptotic variance

By theory, we expect the MCVar to agree closely with the AVar for every one of the estimators in each table. This was confirmed for the majority of the table cells. We computed the ratio (MCVar / AVar). It is shown in parentheses in Table 2. This ratio is near 1 for most of the 25 estimators for a given U_C. The ratio is a bit different from 1 for some cells which are not of primary interest. For example, this occurs when a small U_M is "sandwiched" in the computation between a big U_C and a big U_L, or when a big U_M lies between a small U_C and a small U_L. For these estimators, AVar underestimates the true variance.

11.2. Results for a fixed C-group

For each fixed C-group U_C, Table 1 (which has MCVar) and Table 2 (which has AVar) show the results for the 25 estimators of y_{d+}. Out of these, 20 are indirect (borrowing strength) estimators, namely those in columns $U_L = U_{5000}$, $U_L = U_{2500}$, $U_L = U_{1000}$ and $U_L = U_{600}$. The remaining five estimators, in the column $U_L = U_{500}$, are direct. One of these is the CALWEIGHT, defined by $U_M = U_C$. The five REGFIT estimators are found on the diagonal, that is, when $U_M = U_L$.

Consider the 25 entries in Table 2 for fixed U_C such that $U_d \subset U_C$. That is, the domain is a proper subset of a calibration group. There are four of these groups: $U_C = U_{5000}$, $U_C = U_{2500}$, $U_C = U_{1000}$ and $U_C = U_{600}$. We know from Result 8.1 that CALWEIGHT has the smallest AVar among the 25 values in each of these groups. For example, when $U_C = U_{2500}$, Table 2 shows the minimum value is 940,134. All 20 indirect estimators have a higher AVar than CALWEIGHT, which also has the smallest AVar out of the five direct estimators.

Table 1. Monte Carlo variance of $\hat{y}_{d+}$ *under SRS for* $\mathbf{x}_k = (1, x_k)^T$, $\mathbf{z}_k = \mathbf{z}_{sk} = \mathbf{x}_{Ck} - \bar{\mathbf{x}}_{Cs}$ *and different* U_C, U_M *and* U_L

U_C	U_M	U_L				
		$L = 5{,}000$	$L = 2{,}500$	$L = 1{,}000$	$L = 600$	$L = 500$
$C = 5{,}000$	$M = 5{,}000$	1,485,500	1,158,591	1,046,523	1,040,439	**1,040,074**
	$M = 2{,}500$	2,855,729	1,535,378	1,080,227	1,047,787	1,045,264
	$M = 1{,}000$	15,675,453	5,169,699	1,382,711	1,123,654	1,101,471
	$M = 600$	55,664,447	16,647,895	2,416,787	1,316,962	1,247,356
	$M = 500$	82,813,284	24,399,453	3,097,482	1,453,117	1,331,781
$C = 2{,}500$	$M = 5{,}000$	1,197,931	8,776,427	1,751,298	1,126,672	1,054,162
	$M = 2{,}500$	1,222,620	2,856,874	1,055,159	943,131	**938,217**
	$M = 1{,}000$	2,531,748	18,078,337	2,866,846	1,413,267	1,213,874
	$M = 600$	6,273,725	56,214,333	7,901,223	2,869,123	2,170,815
	$M = 500$	8,739,745	82,733,808	11,506,429	3,974,636	2,882,098
$C = 1{,}000$	$M = 5{,}000$	709,205	3,629,900	8,464,447	2,642,812	1,858,729
	$M = 2{,}500$	1,041,542	1,073,121	2,718,465	981,081	776,454
	$M = 1{,}000$	1,059,066	671,272	1,067,114	634,718	**617,825**
	$M = 600$	1,209,402	1,373,616	3,257,850	1,069,502	825,387
	$M = 500$	1,323,789	2,000,942	4,998,664	1,512,654	1,074,110
$C = 600$	$M = 5{,}000$	541,422	1,765,690	4,910,243	6,049,269	3,844,633
	$M = 2{,}500$	988,997	275,875	917,262	1,224,645	712,752
	$M = 1{,}000$	1,009,973	292,773	273,619	322,187	252,967
	$M = 600$	1,011,596	323,292	248,944	273,415	**244,419**
	$M = 500$	1,018,348	276,515	324,645	401,365	274,668
$C = 500$	$M = 5{,}000$	493,855	1,188,060	3,714,391	4,651,955	4,899,111
	$M = 2{,}500$	972,938	61,103	416,776	627,575	686,746
	$M = 1{,}000$	996,981	195,088	60,321	70,093	75,399
	$M = 600$	999,017	239,006	68,066	60,237	60,739
	$M = 500$	998,976	249,375	71,888	60,678	**60,220**

Table 2. Asymptotic variance of $\hat{y}_{d+}$ under SRS for $\mathbf{x}_k = (1, x_k)^{\mathrm{T}}$, $\mathbf{z}_k = \mathbf{z}_{sk} = \mathbf{x}_{Ck} - \bar{\mathbf{x}}_{Cs}$ and different U_C, U_M and U_L. In parentheses, the ratio of the Monte Carlo variance to the asymptotic variance

U_C	U_M	U_L				
		$L = 5{,}000$	$L = 2{,}500$	$L = 1{,}000$	$L = 600$	$L = 500$
$C = 5{,}000$	$M = 5{,}000$	1,488,121 (0.998)	1,161,159 (0.998)	1,051,469 (0.995)	1,046,232 (0.994)	**1,046,153 (0.994)**
	$M = 2{,}500$	2,820,876 (1.012)	1,537,858 (0.998)	1,080,549 (1.000)	1,049,921 (0.998)	1,048,107 (0.997)
	$M = 1{,}000$	14,536,406 (1.078)	4,931,737 (1.048)	1,382,783 (1.000)	1,106,455 (1.016)	1,085,841 (1.014)
	$M = 600$	48,302,023 (1.152)	14,796,285 (1.125)	2,300,726 (1.050)	1,293,625 (1.018)	1,215,270 (1.026)
	$M = 500$	69,047,159 (1.199)	20,870,854 (1.169)	2,872,564 (1.078)	1,412,687 (1.029)	1,298,258 (1.026)
$C = 2{,}500$	$M = 5{,}000$	1,204,138 (0.995)	8,665,675 (1.013)	1,732,854 (1.011)	1,118,151 (1.008)	1,048,195 (1.006)
	$M = 2{,}500$	1,233,565 (0.991)	2,859,783 (0.999)	1,055,107 (1.000)	944,373 (0.999)	**940,134 (0.998)**
	$M = 1{,}000$	2,505,161 (1.011)	17,769,617 (1.017)	2,869,494 (0.999)	1,402,775 (1.007)	1,204,255 (1.008)
	$M = 600$	6,007,813 (1.044)	54,059,998 (1.040)	7,746,131 (1.020)	2,871,561 (0.999)	2,165,339 (1.003)
	$M = 500$	8,235,994 (1.061)	78,691,698 (1.051)	11,169,500 (1.030)	3,949,164 (1.006)	2,885,773 (0.999)
$C = 1{,}000$	$M = 5{,}000$	706,483 (1.004)	3,397,715 (1.068)	7,861,335 (1.077)	2,403,798 (1.099)	1,689,106 (1.100)
	$M = 2{,}500$	1,047,549 (0.994)	1,076,956 (0.996)	2,728,890 (0.996)	973,577 (1.008)	767,905 (1.011)
	$M = 1{,}000$	1,065,925 (0.994)	673,201 (0.997)	1,071,021 (0.996)	629,835 (1.008)	**612,769 (1.008)**
	$M = 600$	1,206,045 (1.003)	1,343,480 (1.022)	3,204,525 (1.017)	1,072,492 (0.997)	822,131 (1.004)
	$M = 500$	1,311,360 (1.009)	1,935,384 (1.034)	4,876,968 (1.025)	1,508,428 (1.003)	1,077,188 (0.997)
$C = 600$	$M = 5{,}000$	534,239 (1.013)	1,469,525 (1.202)	4,096,244 (1.199)	5,053,992 (1.197)	3,170,367 (1.213)
	$M = 2{,}500$	993,760 (0.995)	274,980 (1.003)	913,322 (1.004)	1,219,820 (1.004)	703,947 (1.013)
	$M = 1{,}000$	1,016,013 (0.994)	294,052 (0.996)	273,300 (1.001)	321,435 (1.002)	249,451 (1.014)
	$M = 600$	1,017,789 (0.994)	324,991 (0.995)	248,983 (1.000)	273,165 (1.001)	**241,565 (1.012)**
	$M = 500$	1,024,122 (0.994)	277,111 (0.998)	321,100 (1.011)	396,590 (1.012)	274,073 (1.002)
$C = 500$	$M = 5{,}000$	483,039 (1.022)	881,015 (1.349)	2,851,184 (1.303)	3,593,203 (1.295)	3,789,620 (1.293)
	$M = 2{,}500$	976,832 (0.996)	58,691 (1.041)	415,671 (1.003)	627,311 (1.000)	686,817 (1.000)
	$M = 1{,}000$	1,002,480 (0.995)	194,371 (1.004)	58,646 (1.029)	68,385 (1.025)	73,709 (1.023)
	$M = 600$	1,004,721 (0.994)	238,751 (1.001)	66,517 (1.023)	58,646 (1.027)	59,134 (1.027)
	$M = 500$	1,004,719 (0.994)	249,240 (1.001)	70,388 (1.021)	59,102 (1.027)	**58,646 (1.027)**

The results on AVar are confirmed by the MCVar in Table 1. It shows CALWEIGHT has the smallest MCVar among the 25 values for every fixed U_C. Thus both AVar and MCVar support a decision to choose CALWEIGHT over other estimators, including the REGFIT estimators. Differences between CALWEIGHT and REGFIT are sometimes large. For example, when $U_C = U_{2500}$, the MCVar of most REGFIT estimators (on the diagonal) is over three times that of CALWEIGHT. For $U_C = U_d = U_{500}$, CALWEIGHT and REGFIT/DOM are the same. We know by Result 8.1 that its AVar is the smallest. Table 2 shows that this AVar has a value of 58,646. The corresponding minimum value in Table 1 is 60,220. Here, the minimum is very flat; other values on the diagonal are very close.

11.3. Comparing results for the five different C-groups

Figure 2 shows the progression of MCVar for CALWEIGHT, REGFIT/SAMPLE and REGFIT/DOM as U_C moves from U_{500} to $U = U_{5000}$. The points for the five U_C groups were joined by spline fitting to produce a curve for the variance of each of the estimators. The figure also shows the variance of the HT estimator as a horizontal line with constant value 1,048,376. For the CALWEIGHT estimator in particular, the value of the auxiliary information diminishes drastically as the calibration group U_C expands away from the fixed domain U_d. The effect is felt as soon as U_C becomes larger than U_d. The MCVar of the optimal CALWEIGHT estimator increases sharply from 60,220 when $U_C = U_d = U_{500}$ to 244,419 when $U_C = U_{600}$. The increase continues but tapers off to a value which is essentially the variance of the HT estimator. The sharp increase in the variance of CALWEIGHT can be explained by saying that the correlation between $\mathbf{x}_{Ck}$ and y_{dk} diminishes due to an increasingly larger proportion of zero values y_{dk}. The gain over the HT estimator is small when we use auxiliary information for a C-group considerably above the domain.

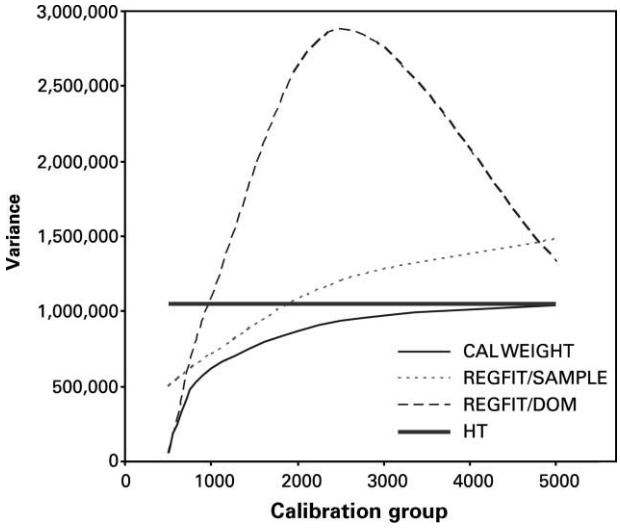

Fig. 2. Comparison of CALWEIGHT, REGFIT/SAMPLE, REGFIT/DOM and HT estimators. The horizontal axis shows the size of the C-group U_C

11.4. The effect of extending the auxiliary vector $\mathbf{x}$ *to include the constant "1"*

We constructed our finite population to have a significant intercept by putting $\alpha = 20$. As a result, when the domain itself is the C-group, $U_C = U_d = U_{500}$, the optimal CALWEIGHT estimator has much smaller variance for $\mathbf{x}_k = (1, x_k)^T$ than for $\mathbf{x}_k = x_k$. However, when the C-group is larger than U_d, there are only minor differences in the variance of the optimal CALWEIGHT estimator between $\mathbf{x}_k = (1, x_k)^T$ and $\mathbf{x}_k = x_k$.

11.5. Comparing with the HT estimator

The HT estimator $\hat{y}_{d\oplus\pi}$ uses no auxiliary information at all. For our population, its variance is 1,048,376. By comparison, although the REGFIT estimators on the diagonal of Table 2 use auxiliary information, they all have a larger MCVar for several of the C-groups. The optimal CALWEIGHT estimator always has smaller MCVar than the HT, but for $U_C = U_{5000}$ it is only slightly smaller at 1,040,074. When $U_C = U_d = U_{500}$, the auxiliary information produces an impressive variance reduction over the HT estimator. The MCVar of CALWEIGHT (= REGFIT/DOM) is then only 6% of that of the HT estimator, even though the correlation between y and x is not exceptionally strong at 0.67.

12. Conclusions and Recommendations

Design-based survey sampling theory is not in all respects like "ordinary statistical theory." This has been evidenced a number of times over the years. A concept in ordinary statistical theory cannot always be transferred to the theory of sampling from finite populations and be expected there to deliver the same effect. An illustration was found in this article: Borrowing strength is not a fruitful concept within the class of design-based domain estimators considered in this article. A direct estimator, the CALWEIGHT estimator, is better than all estimators that borrow strength. The CALWEIGHT estimator is also better than all other direct estimators in the class.

An earlier example of the same kind was Godambe's (1955) proof of the nonexistence of a minimum variance unbiased estimator, within a certain, perfectly viable class of estimators of a finite population. The concept of "minimum variance unbiased estimation" is more limited in its usefulness in design-based sampling theory than in other branches of statistics. Nevertheless, Godambe's result does not contradict Neyman's (1934) optimality results for STSRS, because different classes of estimators were considered by these two authors.

Another example was Godambe's (1966) observation that there exists no unique maximum likelihood estimate of a finite population parameter such as its total or mean. He showed that the likelihood function is constant over "the relevant part" of the N-dimensional parameter space, and zero outside. Because of the "flat likelihood," no unique maximum likelihood estimate is obtained. Thus maximum likelihood is not a fruitful concept in design-based estimation theory, in contrast to its vital role in "ordinary statistical theory."

What is going to be the future role of design-based domain estimation theory? Design-based estimates, such as those presented in this article, are recognized for their objectivity, impartiality and freedom from assumptions. These features make them well suited for the objectives of a statistical agency. On the one hand, some may see it as a weakness that this

branch of survey theory cannot profit from an attractive statistical concept such as borrowing strength. On the other hand, our results caution that any domain of interest can always be so specific that its own y-values resemble no y-values from outside the domain and that, as a consequence, direct estimation prevails.

We believe that our results emphasize the need for clarity in two respects. There is a need for (i) clarity in the presentation made to users of statistical results and the methods behind them; (ii) clarity in norms and guiding principles for the production of domain estimates, particularly in national statistical agencies.

Users differ considerably in their ability to perceive and understand the differences that exist in the interpretation of design-based estimates on the one hand and model-dependent estimates on the other. Still, users do appreciate a clear declaration of the statistical agency's methodological stand in regard to published statistics for domains. If the agency's norm is to publish, whenever possible, design-based domain estimates, then this should be unequivocally declared. The agency owes it to users to explain whether estimates are design-based and thus free of assumptions, or whether the agency has taken the step to produce them via model assumptions and borrowing strength. It should be explained whether published measures of precision for domain estimates are design-based, referring to repeated draws of samples, or are computed from other principles.

Norms or rules for the practice of domain estimation need to be clearly formulated to assist the professionals working in national statistical agencies and similar environments. Expressed norms help to ensure that the production in a statistical agency follows uniform and coherent principles. While norms and rules may be useful in that setting, they cannot be given for scientific activity. Creative research will continue on its own terms. For domain estimation in particular, future research will undoubtedly continue to produce interesting results, both in the theory of design-based domain estimation and in the tradition of model-dependent small area estimation.

13. References

Andersson, C. and Nordberg, L. (1998). CLAN97 - A SAS-program for Computation of Point and Standard Error Estimates in Sample Surveys. Stockholm: Statistics Sweden.

Caron, N., Deville, J.C., and Sautory, O. (1998). Estimation de précision de données issues d'enquêtes: document méthodologique sur le logiciel POULPE. Document de travail de la Direction des Statistiques Démographiques et Sociales No. 9806. INSEE, Paris.

Casady, R.J. and Valliant, R. (1993). Conditional Properties of Post-stratified Estimators under Normal Theory. Survey Methodology, 19, 183–192.

Cochran, W.G. (1977). Sampling Techniques. (3rd edition). New York: Wiley.

Deville, J.C. (2002). La correction de la nonréponse par calage généralisé. Actes des Journeés de Méthodologie, INSEE, Paris.

Deville, J.C. and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Estevao, V.M., Hidiroglou, M.A., and Särndal, C.E. (1995). Methodological Principles for a Generalized Estimation System at Statistics Canada. Journal of Official Statistics, 11, 181–204.

Estevao, V.M. and Särndal, C.E. (1999). The Best Use of Auxiliary Information in Design-based Estimation for Domains. Survey Methodology, 25, 213–221.

Estevao, V.M. and Särndal, C.E. (2000). A Functional Form Approach to Calibration. Journal of Official Statistics, 16, 379–399.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. Statistical Science, 9, 55–93.

Godambe, V.P. (1955). A Unified Theory of Sampling from Finite Populations. Journal of the Royal Statistical Society, Series B, 17, 269–278.

Godambe, V.P. (1966). A New Approach to Sampling from Finite Populations I, II. Journal of the Royal Statistical Society, Series B, 28, 310–328.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Volume 1: Methods and Applications. New York: John Wiley.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-dependent and Probability-sampling Inferences in Sample Surveys (with Discussion). Journal of the American Statistical Association, 78, 776–807.

Kish, L. (1965). Survey Sampling. New York: Wiley.

Marker, D.A. (2001). Producing Small Area Estimates from National Surveys: Methods for Minimizing Use of Indirect Estimators. Survey Methodology, 27, 183–188.

Montanari, G.E. (1987). Post-sampling Efficient Prediction in Large-scale Surveys. International Statistical Review, 55, 191–202.

Montanari, G.E. (1998). On Regression Estimation of Finite Population Mean. Survey Methodology, 24, 69–77.

Montanari, G.E. (2000). Conditioning on Auxiliary Variable Means in Finite Population Inference. Australian and New Zealand Journal of Statistics, 42, 407–421.

Montanari, G.E. and Ranalli, M.G. (2002). Asymptotically Efficient Generalized Regression Estimators. Journal of Official Statistics, 18, 559–589.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society, 97, 558–625.

Nieuwenbroek, N. and Boonstra, H.J. (2002). Bascula 4.0 for Weighting Sample Survey Data with Estimation of Variances. The Survey Statistician – Software Review, July.

Purcell, N.J. and Kish, L. (1979). Estimation for Small Domains. Biometrics, 35, 365–384.

Rao, J.N.K. (1994). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. Journal of Official Statistics, 10, 153–165.

Rao, J.N.K. (1999). Some Recent Advances in Model-based Small Area Estimation. Survey Methodology, 25, 175–186.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Received August 2002
Revised April 2004