METHODS USED IN SAMPLING AND ESTIMATION AT STATISTICS CANADA

M.A. Hidiroglou, R. Carpenter and V. Estevao

Business Survey Methods Division, Statistics Canada

11J, R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A-OT6

Key Words: Sampling, estimation, calibration, calibration factors, two-phase sampling, domains.

1 Introduction

This paper presents the methodology used in the development of Statistics Canada's generalized systems for sampling and estimation. These two systems are known as GSAM and GES.

The methods in GSAM include traditional approaches for stratification, allocation and sampling as well as some new methods for the sampling of periodic surveys. Estimation is based on the concept of calibration to known auxiliary totals. The framework developed for the GES allows the specification of a wide family of calibration estimators for one-stage designs. We also introduce a new proposal for two-phase estimation for future implementation in GES.

2 Sampling

Sampling involves two basic components: sample design and selection. In sample design, the sampling unit is defined and the frame is established. The frame consists of the sampling units belonging to the population of interest. The frame is stratified to produce reliable estimates for variables of interest. The sample size is then determined based on a sample allocation and selection method. Stratification and allocation methods are described in sections 2.1 and 2.2.

Sample selection picks a sample based on the allocation and selection method. In repeated surveys, the sample may be rotated in each survey period to minimize response burden. Sample selection is described in section 2.3. From time to time, the frame is also updated for births, deaths and classification changes.

2.1 Stratification

The units in the frame are stratified using a set of rules that define the strata. A rule is a compound set of logical expressions. The stratum rules are based on categorical or continuous variables on the frame.

In the case of categorical variables, such as geography or industry, stratum rules are defined by creating simple expressions from the unique combinations of their values. For a continuous variable, the rules are created by using either an algorithm or applying a specified partition of the variable values. Examples of algorithms in GSAM include the Dalenius-Hodges cumulative $\sqrt{f}$ rule and a simple univariate clustering procedure that minimizes the mean square error of the variable of interest.

2.2 Allocation

Two methods are used to determine the allocation for single stage designs under simple random sampling without replacement. The first is known as a power allocation. It has been proposed by a number of authors including Bankier (1988). The number of units in each stratum is obtained as the solution to the following minimization problem.

$$\begin{aligned} &\mathrm{Min}\ \textstyle\sum_h (X_h^q\ \mathrm{CV}(\hat{Y}_h))^2 \\ &= \mathrm{Min}\ \textstyle\sum_h \frac{X_h^{2q}}{Y_h^2} N_h^2 (1-f_h) \frac{S_h^2}{n_h} \qquad (2.1) \\ &\text{subject to } \textstyle\sum_h c_h n_h \le c \\ &\text{and } l_h \le n_h \le u_h \end{aligned}$$

This allocation requires information on the variable of interest y and an auxiliary variable x for which the stratum totals X_h are known. There is a constraint based on the fixed overall sample cost (c) and restrictions on the sample size in each stratum. The cost of sampling each unit in stratum h is given by c_h. The upper bound u_h is usually the stratum population size N_h while the lower bound l_h is set based on the required precision for the estimates. The exponent q allows an allocation between two extremes. It provides a Neyman allocation when $q = 1$ and $x = y$. On the other hand, $q = 0$ gives an allocation with approximately equal coefficients of variation for the strata if the ratios $S_h/\overline{Y}_h$ do not vary significantly and the sampling fractions are small.

The other allocation method is motivated by the need to determine the minimum sample cost to meet specific reliability measures in the estimation of selected variables. These constraints are formulated in terms of the coefficient of variation (CV) for the estimated population totals of variables $y_{(1)}, \ldots, y_{(i)}, \ldots, y_{(m)}$. As before, lower and upper bounds may be included in the problem definition. This leads to the following allocation problem.

$$\begin{aligned} &\text{Min } \textstyle\sum_h c_h n_h \\ &\text{subject to } \mathrm{CV}(\hat{Y}_{(i)}) \le v_{(i)} \quad \text{for } i = 1,2,\ldots m \qquad (2.2) \\ &\text{and } l_h \le n_h \le u_h \end{aligned}$$

Each specified upper bound $v_{(i)}$ imposes a constraint on the CV of the estimated total of variable $y_{(i)}$. This constraint may be written as follows.

$$\frac{\sqrt{\sum_h (1-f_h) N_h^2 S_{h(i)}^2 / n_h}}{Y_{(i)}} \le v_{(i)} \qquad (2.3)$$

In both methods, a proxy variable is used in place of $y_{(i)}$ to estimate the stratum population variance $S_{h(i)}^2$ and total $Y_{(i)}$. Therefore, the resulting allocation only provides a good estimate of the required sample sizes. In practice, not all of the CV constraints may be met with the actual sample selected. However, the differences between the realized and required CVs should be small for most samples.

The solution to both of these problems is obtained by an algorithm developed by J. Bethel (1989) and extended by V. Estevao (1993) to include the lower and upper bound constraints.

2.3 *Sample Selection*

The selection method must yield workable selection probabilities for estimation and variance. It must also cope with a changing universe, accommodate rotation to reduce response burden and handle classification changes. The methods that were considered included Bernoulli sampling, rotation groups and collocated sampling.

For Bernoulli sampling, Brewer (1972) introduced the idea of assigning a selection number to each unit. The selection number is simply a random number from the uniform distribution $U(0,1)$. Those units with a selection number in the sampling interval or window $[0, f_h]$ are in the sample. The advantage of this procedure is its simplicity. However, the sample size in each stratum is random and this can be a problem for small strata.

The rotation group method, given by Hidiroglou, Choudry and Lavalée (1991), distributes the units in stratum h into P_h rotation groups. Each rotation group gets an assignment order and each unit is assigned to a rotation group using this order. Those units in the first p_h panels are in the sample where $p_h = \lfloor f_h \times P_h \rfloor$ and $\lfloor x \rfloor$ is the integer portion of x. Both P_h and p_h are calculated from the sampling fraction f_h, the length of time a sampled unit stays in the sample $t(in)_h$ and the amount of time it remains out $t(out)_h$.

Collocated sampling is similar to Bernoulli but it includes one additional step. The units of each stratum are randomly assigned selection numbers equally spaced on the interval $[0,1]$ according to the following formula.

$$v_i = \frac{\mathrm{rank}(u_i) - \delta_h}{N_h} \quad \text{for } i \text{ in stratum } h \qquad (2.4)$$

where u_i, δ_h are random values from $U(0,1)$,
rank(u_i) is the order of u_i among $u_1, u_2, \ldots, u_{B_h}$
and N_h is the number of units in stratum h.

The value of δ_h is used to provide a random start in the distribution of the units on the $[0,1]$ interval.

In subsequent periods, new units or births are assigned selection numbers by applying (2.4) to this group of units, with the number of births in the stratum replacing N_h.

Two common features of the above procedures are:

(i) the use of a sampling interval to determine which units are selected; and

(ii) rotation of the sampling units.

These features are explained for the specific case of collocated sampling. Initially, the sampling window is $[0, f_h]$ where f_h is the sampling fraction. Using a fixed sampling fraction for selecting the sample does not always result in a fixed sample size for repeated survey periods because of changes to the frame over time.

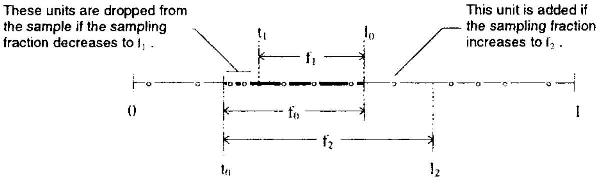
For a repeated survey, sample rotation may be used to reduce response burden. Sample rotation is performed by simply moving the sampling window to allow new units to replace part of the sample. The amount moved is given by the rotation increment r_h using the following formula.

$$r_h = \min\left[\frac{f_h}{t(in)_h},\ \frac{(1-f_h)}{t(out)_h}\right] \qquad (2.5)$$

Here, $t(in)_h$ is the time that a unit should remain in the sample and $t(out)_h$ is the minimum time it should be out of the sample once it has been rotated out. Both requirements are met if $r_h = \frac{f_h}{t(\text{in})_h}$. Otherwise, the time in sample exceeds $t(in)_h$.

With the addition of new units, the removal of units and changes to the sampling parameters and stratification, the selection numbers may no longer be equally spaced on $[0,1]$. As a result, there may be a large difference between the required sampling fraction and the observed sampling fraction calculated from the actual number of units in the sample. Rebalancing reestablishes equally spaced selection numbers and aligns them relative to the current sampling window to retain as much of the sample as possible within the sampling window. The rebalance methodology is an extension of the work done by R. Carpenter (1990).

This approach is based on a simple rationale for minimizing response burden. If the sampling fraction has increased, the oldest units not in the sample are the first to be added into the sample. If the sampling fraction has decreased, the oldest sample units are removed. The easiest way to see this is to consider the simple case of a stratum where the population has not changed from the previous period but the sampling fraction changes between periods. Suppose the previous sampling fraction is f_0 and the sampling window is $[\mathrm{t}_0, \mathrm{l}_0]$. The following diagram shows the position of the new sampling window if f_0 either decreases to f_1 or increases to f_2.


These units are dropped from the sample if the sampling fraction decreases to f_1.
t_1
l_0
This unit is added if the sampling fraction increases to f_2.
f_1
0
f_0
1
f_2
t_0
l_2

After rebalancing, the sample size (the number of units in the sampling window) is $n_h = \lfloor f_h \times N_h + u_h \rfloor$, where u_h is $U(0,1)$. The observed sampling fraction is as close as possible to the required sampling fraction f_h. The difference between f_h and the observed sampling fraction n_h/N_h always satisfies the following condition.

$$0 \leq \left| f_h - \frac{n_h}{N_h} \right| < \frac{1}{N_h} \qquad (2.6)$$

3 Estimation for One-Stage Designs

The Generalized Estimation System uses auxiliary information to produce estimates for one-stage designs. The underlying theory was developed by Särndal, Swenson, and Wretman (1989) and the framework for this implementation is described in Estevao, Hidiroglou, and Särndal (1995). This theory is based on regression estimators known as GREG (Generalized Regression Estimator) and their extension to the wider family of calibration estimators. Most estimators used in survey practice, including the post-stratified and raking ratio estimators, are special members of the GREG. In this paper we present one-stage designs. The extension to multi-stage designs has been given by Estevao, Hidiroglou, and Särndal (1995).

3.1 One-Stage Element Sampling

Consider the estimation of the total Y of variable y over the population of elements given by $U = \{1,\ldots,k,\ldots,N\}$. A probability sample s is selected from U with inclusion probability π_k for $k \in s$. These provide the sample design weights $w_k = 1/\pi_k$ for $k \in s$. Without the use of any additional information, the estimated total of Y is given by the π-estimator below.

$$\hat{Y}_\pi = \sum_s w_k y_k \qquad (3.1)$$

Suppose the following information is known for a set of auxiliary variables given by $\boldsymbol{x}_k' = (x_{1k},\ldots,x_{jk},\ldots x_{Jk})$ for $k \in s$.

$$\{\boldsymbol{x}_k\} \text{ for } k \in U \text{ or } \boldsymbol{X} = \sum_U \boldsymbol{x}_k \qquad (3.2)$$

This auxiliary information is used to obtain new sample weights $\tilde{w}_k = w_k g_k$ under the calibration approach proposed by Deville and Särndal (1992). The multiplicative adjustment g_k is known as a g-weight or calibration factor. The idea behind this approach is to find a set of weights $\tilde{w}_k$ that satisfy the calibration equations given below.

$$\sum_s \tilde{w}_k \boldsymbol{x}_k = \sum_U \boldsymbol{x}_k \qquad (3.3)$$

These equations are simply a restatement of the requirement that the estimates for the auxiliary totals equal the known totals. In general, there are many sets of weights that satisfy these equations. The optimal set of weights $\tilde{w}_k$ should be close to the original design weights in order to retain the sampling properties of the sample design. This means that the values of g_k should be close to 1 for most elements in the sample.

A least squares distance function is used to provide a measure of the distance between w_k and the $\tilde{w}_k$. The calibration problem can be stated as follows.

$$\begin{aligned} &\text{Min } \sum\nolimits_s \frac{c_k(\tilde{w}_k - w_k)^2}{2w_k} \\ &\text{subject to } \sum\nolimits_s \tilde{w}_k \boldsymbol{x}_k = \sum\nolimits_U \boldsymbol{x}_k \end{aligned} \qquad (3.4)$$

The positive values c_k provides a more general weighting of the individual terms of the distance measure. Many of the traditional estimators are obtained by assuming $c_k = 1$. With a single positive auxiliary variable x_k, putting $c_k = x_k$ leads to the simple ratio estimator when $x_k > 0$.

The solution to the calibration problem can be written for each $k \in s$ as follows.

$$\begin{aligned} &\tilde{w}_k = w_k\{1+(\sum\nolimits_U \boldsymbol{x}_k - \sum\nolimits_s w_k \boldsymbol{x}_k)' \boldsymbol{T}^{-1} \boldsymbol{x}_k / c_k\} \\ &\quad \text{where } \boldsymbol{T} = (\sum\nolimits_s w_k \boldsymbol{x}_k \boldsymbol{x}_k' / c_k) \end{aligned} \qquad (3.5)$$

In terms of the calibration factors g_k, we have the following for each $k \in s$.

$$g_k = 1+(\sum\nolimits_U \boldsymbol{x}_k - \sum\nolimits_s w_k \boldsymbol{x}_k)' \boldsymbol{T}^{-1} \boldsymbol{x}_k / c_k \qquad (3.6)$$

Using the weights $\tilde{w}_k$ from the calibration, the GREG estimator of Y is then given by the expression.

$$\hat{Y} = \sum\nolimits_s \tilde{w}_k y_k \qquad (3.7)$$

It is easy to show that the simple expansion and ratio estimators are obtained under the following specifications for $\boldsymbol{x}_k$ and c_k

	$\boldsymbol{x}_k$	c_k	g_k	$\hat{Y}$
Expansion	1	1	$\frac{N}{\hat{N}_\pi}$	$\frac{N}{\hat{N}_\pi}\hat{Y}_\pi$
Ratio	x_k	x_k	$\frac{X}{\hat{X}_\pi}$	$\frac{X}{\hat{X}_\pi}\hat{Y}_\pi$

The variance of $\hat{Y}$ depends on the residuals from the linear regression of y_k on $\boldsymbol{x}_k$. If the residual variation is small, then $\hat{Y}$ is a more efficient estimator than $\hat{Y}_\pi$. This is true whether or not the linear regression provides an appropriate fit. An estimate of the variance of $\hat{Y}$ can be obtained by the method proposed in Särndal, Swensson, and Wretman (1992).

$$v(\hat{Y}) = \sum\nolimits_{k\in s} \sum\nolimits_{l \in s} w_{kl}(w_k w_l - 1)(g_k e_k)(g_l e_l) \qquad (3.8)$$

$$\text{where } w_{kl} = (\pi_{kl})^{-1} \text{ with } \pi_{kl} = P(k \text{ and } l \in s) \qquad (3.9)$$

$$\begin{aligned} \text{and } e_k &= y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}} \\ &= y_k - \boldsymbol{x}_k' \boldsymbol{T}^{-1} (\sum\nolimits_s w_k \boldsymbol{x}_k y_k / c_k) \end{aligned}$$

The concept of calibration based on totals for the entire population can be extended to individual groups of the population such as strata or post-strata. In either of these cases, we have a partition of the population into mutually exclusive and exhaustive groups U_1, ..., U_p,... U_P. Let us assume the following auxiliary information is available within each group.

$$\text{(i)} \quad \boldsymbol{x}_k' = (x_{1k}, \ldots, x_{jk}, \ldots x_{Jk}) \text{ for } k \in s_p \qquad (3.10)$$

$$\text{(ii)} \quad \boldsymbol{X}_p = \sum\nolimits_{U_p} \boldsymbol{x}_k$$

It is possible to carry out the calibration within each group and obtain a new set of weights $\tilde{w}_k$ for estimation. For each $k \in s_p$ we have the following.

$$\begin{aligned} &\tilde{w}_k = w_k\{1+(\sum\nolimits_{U_p} \boldsymbol{x}_k - \sum\nolimits_{s_p} w_k \boldsymbol{x}_k)' \boldsymbol{T}_p^{-1} \boldsymbol{x}_k / c_k\} \\ &\quad \text{where } \boldsymbol{T}_p = (\sum\nolimits_{s_p} w_k \boldsymbol{x}_k \boldsymbol{x}_k' / c_k) \end{aligned} \qquad (3.11)$$

These weights are not the same as those obtained by calibration over the entire population. The two approaches lead to two familiar types of estimators.

When we partition the population into mutually exclusive and exhaustive groups, we obtain a separate estimator. When the population is the only calibration group, we get a combined estimator. For the ratio estimator, we have the following results.

	$\hat{Y}$
Separate Ratio	$\sum\nolimits_p \frac{X_p}{\hat{X}_{\pi p}} \hat{Y}_{\pi p}$
Combined Ratio	$\frac{X}{\hat{X}_\pi} \hat{Y}_\pi$

The only difference between a stratified and a post-stratified estimator is in the definition of the calibration groups. Stratified estimators are obtained when the calibration groups correspond to the strata or groups of strata. Post-stratified estimators are produced when the calibration groups are other than strata or groups of strata. Note the similarity of the formulas for the stratified and post-stratified ratio estimators.

	$\hat{Y}$
Stratified Ratio	$\sum_h \frac{X_h}{\hat{X}_{\pi h}} \hat{Y}_{\pi h}$
Post-stratified Ratio	$\sum_p \frac{X_p}{\hat{X}_{\pi p}} \hat{Y}_{\pi p}$

More general estimators are obtained by an arbitrary partition of the population into calibration groups or specifying different auxiliary variables in each group or setting arbitrary values for c_k The only requirement is that we must know the auxiliary totals for the variables in each of the defined groups.

3.2 *One-Stage Cluster Sampling*

Under one-stage cluster sampling, a sample s of clusters is selected and all elements in these clusters form the sample of elements. For each cluster $i \in s$ and element $k \in i$ we have $w_i = w_k$ where $w_i = 1/\pi_i$. For this type of design, it is possible to have auxiliary information for the elements or the clusters. The calibration concepts can be extended as follows.

If auxiliary information is available for the clusters, we can partition the clusters into calibration groups U_p for which we know $\boldsymbol{x}_i$ for $i \in s_p$ and $\sum_{U_p} \boldsymbol{x}_i$. Then we find the cluster weights w_i within each group that solve the calibration problem on the cluster auxiliary data.

$$\begin{aligned} &\text{Min} \sum_{s_p} \frac{c_i(\tilde{w}_i - w_i)^2}{2w_i} \\ &\text{subject to} \sum_{s_p} \tilde{w}_i \boldsymbol{x}_i = \sum_{U_p} \boldsymbol{x}_i \end{aligned} \tag{3.12}$$

This leads to the following estimator of the total of y over the population of elements.

$$\hat{Y} = \sum_s \tilde{w}_i \sum_{k \in i} y_k = \sum_s \tilde{w}_i Y_i \tag{3.13}$$

The variance of this estimator can be obtained from section 3.1 with residuals given as follows.

$$\begin{aligned} e_i &= Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{B}} \\ &= Y_i - \boldsymbol{x}_i'(\sum_s w_i \boldsymbol{x}_i \boldsymbol{x}_i / c_i)^{-1} (\sum_s w_i \boldsymbol{x}_i Y_i / c_i) \end{aligned} \tag{3.14}$$

A similar approach is used when the auxiliary information is known for calibration groups formed by a partition of the elements. This leads to the formulas shown in section 3.1.

It is important to note that these two approaches are based on partitions of different populations. A partition of the elements allows the elements of a cluster to be assigned to different calibration groups. Even when the groups are the individual strata, the difference in the level of the auxiliary information leads to different estimators. Consider the following example with strata as the calibration groups for the elements and the clusters.

Clusters	$\boldsymbol{x}_i = 1$	$c_i = 1$	$\sum_{U_h} \boldsymbol{x}_i = N_h$	$\hat{Y} = \sum_h \frac{N_h}{\hat{N}_{\pi h}} \hat{Y}_\pi$
Elements	$\boldsymbol{x}_k = 1$	$c_k = 1$	$\sum_{U_h} \boldsymbol{x}_k = M_h$	$\hat{Y} = \sum_h \frac{M_h}{\hat{M}_{\pi h}} \hat{Y}_\pi$

In this example, we have the same calibration groups and definition of auxiliary variable. The known totals N_h and M_h represent the number of clusters and elements within the stratum population. With the auxiliary data at the cluster level we obtain the separate expansion estimator. At the element level, we get the ratio to size estimator.

4 *Estimation for Two-Phase Designs*

Two-phase sampling as pointed out in Cochran (1977) is a powerful and cost-effective technique. The incorporation of available data at the population and first phase levels in the estimation process usually yields substantial reductions in the variance of the estimates. The gain will depend on the correlation between the auxiliary data and the variables of interest. Some of the uses of auxiliary data in this manner at Statistics Canada are given in Armstrong and St. Jean (1993), and Hidiroglou et. al (1995). The specific estimation procedures in these papers have been recently generalized by Hidiroglou and Särndal (1995), providing a unified theory for two-phase sampling with auxiliary information. This general theory for two-phase sampling is easily amenable to programming and it will be eventually incorporated into Statistics Canada's Generalized Estimation System. As in Section 3, the approach is via calibration.

A first phase probability sample s_1 ($s_1 \subseteq U$) is drawn from the population $U = \{1, \ldots, k, \ldots, N\}$ such that each unit k has probability π_{1k} of inclusion in the sample. Given that s_1 has been drawn, the second-phase sample s_2 ($s_2 \subseteq s_1 \subseteq U$) is selected from s_1, with selection probabilities $\pi_{k|s_1}$. The first-phase sampling weight of unit k is denoted as $w_{1k} = 1/\pi_{1k}$ and the second phase sampling weight is $w_{2k} = 1/\pi_{k|s_1}$. The overall sampling weight for a sampled unit $k \in s_2$ is $w_k^* = w_{1k} w_{2k}$.

Assuming that the full data auxiliary vector is $\boldsymbol{x}_k$, we decompose it as $\boldsymbol{x}_k = (\boldsymbol{x}'_{1k}, \boldsymbol{x}'_{2k})'$. Here, $\boldsymbol{x}_{1k}$ is a vector for which information is available up to the full population level, and $\boldsymbol{x}_{2k}$ is a vector for which information is available up to the level of the first sample only. Both types of information are important. The following table summarizes our assumptions on the auxiliary information available for estimation.

Set of units	Available Data
Population	$\{\boldsymbol{x}_{1k}\}$ for $k \in U$ or $\sum_U \boldsymbol{x}_{1k}$
First phase sample	$\{(\boldsymbol{x}_{1k}, \boldsymbol{x}_{2k})\}$ for $k \in s_1$
Second phase sample	$\{(\boldsymbol{x}_{1k}, \boldsymbol{x}_{2k}, y_k)\}$ for $k \in s_2$

Given that the design weights are w_k^*, we seek a set of weights $\tilde{w}_k^*$ that lie as close to them as possible. These weights can be obtained through two successive stages of calibration using the function given in section 3.

The first phase calibration factors are denoted as g_{1k}, while the second phase calibration factors are given by g_{2k}. The calibration with respect to both phases produces overall calibration factors $g_k^* = g_{1k} g_{2k}$ for $k \in s_2$. As a result we have: (i) first phase calibrated weights $\tilde{w}_{1k} = w_{1k} g_{1k}$ for $k \in s_1$; (ii) overall calibrated weights $\tilde{w}_k^* = w_k^* g_k^*$ for $k \in s_2$, where $w_k^* = w_{1k} w_{2k}$ is the overall sampling weight.

<u>First phase calibration</u> (from s_1 to U).

Use the first phase sampling weights $\{w_{1k} : k \in s_1\}$ as starting weights. Let $\{c_{1k} : k \in s_1\}$ be specified positive weights. Determine first phase calibrated weights $\tilde{w}_{1k}$ by minimizing

$$\sum_{s_1} \frac{c_{1k}(\tilde{w}_{1k} - w_{1k})^2}{2w_{1k}} \qquad (4.1)$$

subject to

$$\sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_{1k} = \sum_U \boldsymbol{x}_{1k} \qquad (4.2)$$

where the total $\sum_U \boldsymbol{x}_{1k}$ is assumed to be known. Since the total $\sum_U \boldsymbol{x}_{2k}$ is not known at the population level, it is not part of the restrictions. The calibrated weights are $\tilde{w}_{1k} = w_{1k} g_{1k}$ for $k \in s_1$ with

$$g_{1k} = 1 + (\sum_U \boldsymbol{x}_{1k} - \sum_{s_1} w_{1k} \boldsymbol{x}_{1k})' \boldsymbol{T}_1^{-1} \boldsymbol{x}_{1k} / c_{1k} \qquad (4.3)$$

where

$$\boldsymbol{T}_1 = \sum_{s_1} \frac{w_{1k} \boldsymbol{x}_{1k} \boldsymbol{x}'_{1k}}{c_{1k}} \qquad (4.4)$$

<u>Second phase calibration</u> (from s_2 to s_1).

We use as starting weights $\{\tilde{w}_{1k} w_{2k} : k \in s_2\}$. The second phase calibration improves the weights by including information available from the first phase sample. Depending on the specification of the calibration the overall calibration factors g_k^* can be expressed either multiplicatively as the product of the first phase and second phase factors, or additively as a linear combination of these factors. These two formulations correspond to two different GLS distance functions.

<u>Case A</u> (*Multiplicative g-factors*): Starting with the weights $\tilde{w}_{1k} w_{2k}$, determine the overall calibrated weights $\tilde{w}_k^*$ by minimizing

$$\sum_{s_2} \frac{c_{2k}(\tilde{w}_k^* - \tilde{w}_{1k} w_{2k})^2}{2\tilde{w}_{1k} w_{2k}} \qquad (4.5)$$

subject to the second phase calibration equations

$$\sum_{s_2} \tilde{w}_k^* \boldsymbol{x}_k = \sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_k \qquad (4.6)$$

where $\{c_{2k} : k \in s_2\}$ are specified positive weights and $\boldsymbol{x}_k = (\boldsymbol{x}'_{1k}, \boldsymbol{x}'_{2k})'$. The weights resulting from this calibration define the overall calibrated weights. They are given for $k \in s_2$ as

$$\tilde{w}_k^* = \tilde{w}_{1k} w_{2k} g_{2k}^M = w_{1k}^* g_{1k} g_{2k}^M \qquad (4.7)$$

where

$$g_{2k}^M = 1 + (\sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_k - \sum_{s_2} \tilde{w}_{1k} w_{2k} \boldsymbol{x}_k)' (\boldsymbol{T}_2^M)^{-1} \boldsymbol{x}_k / c_{2k} \qquad (4.8)$$

and

$$\boldsymbol{T}_2^M = \sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} \boldsymbol{x}_k \boldsymbol{x}'_k}{c_{2k}} \qquad (4.9)$$

The calibration given by (4.6) assures that the first and second phase estimates of the unknown population total of $\boldsymbol{x}_2$ agree. The overall calibration factor is $g_k^* = g_{1k} g_{2k}^M$.

The factors $c_{2k}/\tilde{w}_{1k}w_{2k}$ are not necessarily all positive. This is because g_{1k} can be zero or negative. The following modification to distance function (4.5) eliminates this problem.

Case B (*Additive g-factors*): An alternative for the second phase calibration is to replace (4.5) by

$$\sum_{s_2} \frac{c_{2k}(\tilde{w}_k^* - \tilde{w}_{1k}w_{2k})^2}{2w_k^*} \tag{4.10}$$

where $\{c_{2k}: k \in s_2\}$ are specified positive weights. Note that the factors c_{2k}/w_k^* are always positive. The overall calibrated weights resulting from minimizing (4.10) subject to (4.6) is

$$\tilde{w}_k^* = w_k^*(g_{1k} + g_{2k}^A - 1) \tag{4.11}$$

where

$$g_{2k}^A = 1 + (\sum_{s_1} \tilde{w}_{1k}\boldsymbol{x}_k - \sum_{s_2} \tilde{w}_{1k}w_{2k}\boldsymbol{x}_k)'(\boldsymbol{T}_2^A)^{-1}\boldsymbol{x}_k \,/\, c_{2k} \tag{4.12}$$

for $k \in s_2$ with

$$\boldsymbol{T}_2^A = \sum_{s_2} \frac{w_k^*\boldsymbol{x}_k\boldsymbol{x}_k'}{c_{2k}} \tag{4.13}$$

This yields the overall calibration factor $g_k^* = g_{1k} + g_{2k}^A - 1$.

Summarizing Cases A and B, the overall calibrated weights are $\tilde{w}_k = w_k^* g_k^*$ where

$$g_k^* = \begin{cases} g_{1k} + g_{2k}^A - 1 & \text{for Case A} \\ g_{1k}\, g_{2k}^M & \text{for Case B} \end{cases} \tag{4.14}$$

Comparing the expressions for g_{2k}^M and g_{2k}^A, we note that the only difference between them lies in the weighting applied in the matrices $\boldsymbol{T}_2^M$ and $\boldsymbol{T}_2^A$.

The final calibrated weights are then given by $\tilde{w}_k^* = w_k^* g_k^*$, where $w_k^* = w_{1k}\, w_{2k}$ is the product of the first and second phase sampling weights. The resulting estimator of Y that incorporates the two levels of auxiliary information is given by

$$\hat{Y} = \sum_{s_2} \tilde{w}_k^* y_k \tag{4.15}$$

As in Section 3, it is possible to extend this theory to incorporate auxiliary information for mutually exclusive and exhaustive calibration groups at the population level and at the level of the first phase sample where these two partitions may be quite different. The details are provided in Hidiroglou and Särndal (1995).

The variance estimator of the two-phase regression estimator $\hat{Y}$ is given in Särndal, Swensson, and Wretman (1992). It is calculated as a total of two components, one for each phase, according to the following formula

$$\begin{aligned} v(\hat{Y}) = {} & \sum_{k \in s_2} \sum_{l \in s_2} w_{2kl}(w_{1k}w_{1l} - w_{1kl})(g_{1k}e_{1k})(g_{1l}e_{1l}) \\ & + \sum_{k \in s_2} \sum_{l \in s_2} w_{1k}w_{1l}(w_{2k}w_{2l} - w_{2kl})(g_{2k}e_{2k})(g_{2l}e_{2l}) \end{aligned} \tag{4.16}$$

where the weights $w_{1k} = 1/\pi_{1k}$ and $w_{1kl} = 1/\pi_{1kl}$ with $\pi_{1kl} = P(k \text{ and } l \in s_1)$ are associated with the first phase of sampling, and $w_{2k} = 1/\pi_{2k}$ and $w_{2kl} = 1/\pi_{2kl}$ with $\pi_{2kl} = P(k \text{ and } l \in s_2 | s_1)$ are their respective counterparts for the second phase. Note that for $k = l$, we have $w_{1kl} = w_{1k}$ and $w_{2kl} = w_{2k}$ in (4.16).

The two sets of residuals required for this variance estimator are as follows.

$$e_{1k} = y_k - \boldsymbol{x}_{1k}'\hat{\boldsymbol{B}}_1 \text{ for } k \in s_1 \tag{4.17}$$

and

$$e_{2k} = y_k - \boldsymbol{x}_k'\hat{\boldsymbol{B}}_2 \text{ for } k \in s_2 \tag{4.18}$$

where

$$\hat{\boldsymbol{B}}_1 = T_1^{-1}\left\{\sum_{s_1} \frac{w_{1k}\boldsymbol{x}_{1k}\hat{y}_{2k}}{c_{1k}} + \sum_{s_2} \frac{w_k^*\boldsymbol{x}_{1k}(y_k - \hat{y}_{2k})}{c_{1k}}\right\} \tag{4.19}$$

$$\text{with } \hat{y}_{2k} = \boldsymbol{x}_k'\hat{\boldsymbol{B}}_2 \text{ and } \hat{\boldsymbol{B}}_2 = T_2^{-1}\sum_{s_2} \frac{w_k^*\delta_k\boldsymbol{x}_k y_k}{c_{2k}} \tag{4.20}$$

$$\text{where } \delta_k = \begin{cases} 1 & \text{for Case A} \\ g_{1k} & \text{for Case B} \end{cases}. \tag{4.21}$$

5. *Domain estimation*

The ideas in sections 3 and 4 are easily carried over to the estimation for domains. Let U_d $(U_d \subseteq U)$ denote a domain of U. The total of variable y in U_d may be written as $Y(d) = \sum_{U_d} y_k = \sum_U y(d)$ where $y_k(d)$ is defined as follows.

$$y_k(d) = \begin{cases} y_k & \text{if } k \in U_d \\ 0 & \text{if } k \notin U_d \end{cases} \tag{5.1}$$

The calibrated weights can then be used to produce an estimate $\hat{Y}(d)$ for the domain total based on the observed sample. For the two-phase design of section 4, we use the formula given by (4.15) to produce the estimator of $Y(d)$ as follows.

$$\hat{Y}(d)=\sum\nolimits_{s_2}\tilde{w}_k^* y_k(d) \qquad (5.2)$$

The variance of estimator (5.2) is obtained by formula (4.16), provided y_k is replaced throughout the calculations with $y_k(d)$. This means that the residuals in the formula, e_{1k} and e_{2k}, become the following.

$$e_{1k}(d)=y_k(d)-\boldsymbol{x}'_{1k}\hat{\boldsymbol{B}}_1(d) \text{ for } k\in s_1 \qquad (5.3)$$

and

$$e_{2k}(d)=y_k(d)-\boldsymbol{x}'_k\hat{\boldsymbol{B}}_2(d) \text{ for } k\in s_2 \qquad (5.4)$$

where $\hat{\boldsymbol{B}}_1(d)$ and $\hat{\boldsymbol{B}}_2(d)$ are calculated from the expressions (4.19) and (4.20) for $\hat{\boldsymbol{B}}_1$ and $\hat{\boldsymbol{B}}_2$ by replacing y_k by $y_k(d)$.

6. Summary

The methods provided in this paper are the basis for the development of the Generalized Sampling and Estimation Systems. We have attempted to build a general framework for these systems by using current and relevant methodologies and implementing these through a modular approach. The calibration framework of the GES is the result of this kind of generalized approach to development. However, there is much work to be done to build on this foundation.

BIBLIOGRAPHY

Armstrong, J. and St-Jean, H. (1993). Generalized Regression Estimation for a Two-Phase Sample of Tax Records. *Survey Methodology*, Vol. 20, pp. 91-105.

Bankier, M. (1988). Power Allocations: Determining Sample Sizes for Subnational Areas. *The American Statistician*, Vol. 42, No. 3, pp. 174-177.

Bethel, J. W.. (1989). Sample Allocation in Multivariate Surveys. *Survey Methodology*, Vol. 15, No. 1, pp. 47-57.

Binder, D.A. (1996). Linearization Methods for Single Phase and Two Phase Samples: A Cookbook approach. To appear, *Survey Methodology*.

Carpenter, R.M. (1990). *Resampling Methodology for Rotating Samples*. Statistics Canada, internal report, January 1990.

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

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

Deville, J.-C., Särndal, C.E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, Vol. 88, pp. 1013-1020.

Estevao, V., (1993). Optimum Allocation for Stratified One-Stage SRSWOR Designs. Statistics Canada, internal report, May 1993.

Estevao, V., Hidiroglou, M.A., and Särndal, C.E. (1995). Methodological Principles for a Generalized Estimation System at Statistics Canada. *Journal of Official Statistics*, Vol. 11, No. 2, pp. 181-204.

Hidiroglou, M.A., and Särndal, C.E. (1995). Use of Auxiliary Information for Two-phase Sampling, unpublished paper.

Hidiroglou, M.A., Choudry G.H. and Lavallée, P. (1991). A Sampling and Estimation Methodology for Sub-Annual Busineess Surveys. *Survey Methodology*, Vol. 17, No. 2, pp. 195-210.

Hidiroglou, M.A., and Lavallée, P. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, Vol. 14, No. 1, pp. 33-43.

Hidiroglou, M.A., Latouche, M., Armstrong, B., and Gossen, M. (1995). Improving Survey Information Using Administrative Records: The Case of the Canadian Employment Survey. *Proceedings, Annual Research Conference, U.S. Bureau of the Census*, pp. 171-197.

Särndal, C.E., Swensson, B., and Wretman, J.H. (1989). *The Weighted Residual Technique for Estimating the Variance of the Generalized Regression Estimator of the Finite Population Total.* Biometrika, Vol. 76, No. 3, pp. 527-537.

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