Use of Auxiliary Information for Two-phase sampling

M.A. Hidiroglou and C.E. Särndal

M.A. Hidiroglou, Business Survey Methods Division, Statistics Canada, 11J, R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A-OT6

KEY WORDS: Generalized regression, two-phase sampling, model assisted approach, domains, calibration factors.

1. Introduction

Two-phase sampling is a powerful and cost-effective technique. Rao (1973) applied it to stratification, non response problems and analytic comparisons. Cochran (1977) provided some basic results for two-phase sampling in his book. Current related work in this area includes Breidt and Fuller (1993), Chaudhuri and Roy (1994), and Dupont (1995). Breidt and Fuller (1993) gave numerically efficient estimation procedures for three-phase sampling, in the presence of auxiliary information. Chaudhuri and Roy (1994) studied the optimal properties of well known regression estimators used in two-phase sampling. Särndal and Swensson (1987) provide a framework for regression estimation in two-phase sampling. Dupont (1995) provides additional regression estimators that can be viewed as alternatives to the procedures proposed in this paper. The use of two-phase sampling with administrative files requires some further extensions of the current knowledge in this area. Current efforts to extend the Generalized Estimation System at Statistics Canada require more unified and systematized sampling theory for two-phase designs. The current work is a further step in this direction.

This paper provides some general theory for two-phase sampling for domain estimation. We allow arbitrary sampling designs at both phases of sampling. Furthermore, auxiliary information can be used at either phase of sampling. The auxiliary information comes as known auxiliary variable totals. This auxiliary information is incorporated in the estimation process by calibrating procedures or via regression fitting in each phase. The resulting estimator and its estimated variance can then be expressed in terms of; (I) the original sampling weights; (ii) the calibration factors that reflect the auxiliary data; and (iii) the observed data on the variable of interest. Variances are estimated via the Taylor expansion procedure.

The paper is organized as follows. Section 2 sets up the notation. Section 3 explains how calibration is carried out in each of the two phases using a generalized least squares distance. Section 4 points out how the same estimators can be derived from a regression fit at each stage. Estimated variances for two-phase regression are provided in Section 5. Section 6 explains how the theory applies to domain estimation. Section 7 extends the preceding material through the concept of calibration groups. Section 8 provides a summary.

3 Notation

The population is represented by U = { 1, . . . , k, . . . , N }. A first phase probability sample s_1 ($s_1 \subseteq U$) is drawn from the population U, using a sampling design that generates the selection probabilities π_{1k}. 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 a sampling design with the selection probabilities $\pi_{2k} = \pi_{k|s_1}$. Note the conditional nature of the second phase selection probabilities. From this point on, we work with weights in the estimation process. The first-phase sampling weight of unit k will be denoted as $w_{1k} = 1/\pi_{1k}$, and the second phase sampling weight as $w_{2k} = 1/\pi_{k|s_1}$. The overall sampling weight for a selected unit $k \in s_2$ will be $w_k^* = w_{1k} w_{2k}$.

We next introduce the auxiliary information. Our general notation for the auxiliary vector is x and its value for the k[th] unit is denoted as x_k. As in Särndal, Swensson and Wretman (1992, chapter 9), we partition x_k as $x_k = (x_{1k}', x_{2k}')'$. Here, x_{1k} is a vector for which information is available up to the full population level, and 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. More precisely, we assume that:

(I) x_{1k} is known for all units $k \in U$, or that $\sum_U x_{1k}$ is known and x_{1k} observed for all $k \in s_1$;

(ii) x_{2k} is observed for all $k \in s_1$;

(iii) y_k is observed for all $k \in s_2$.

The following table summarizes our assumptions on the auxiliary information available for estimation.

Table 1: Relationships between set of units and available data at different levels

Set of units	Data available
Population	$\{x_{1k}: k \in U\}$ or $\Sigma_U x_{1k}$
First phase sample	$\{(x_{1k}, x_{2k}): k \in s_1\}$
Second phase sample	$\{(x_{1k}, x_{2k}, y_k): k \in s_2\}$

As we will see in the next section, the presence of auxiliary information will produce calibration factors to be used in the estimation process. The first phase calibration factors are denoted as g_{1k}, while the second phase yields calibration factors denoted as g_{2k}. The calibration with respect to both phases produces overall calibration factors denoted as g_k^*. As a result we will 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. Depending on the specification of the calibration, the g_k^* can be expressed either multiplicatively as the product of the first phase and second phase g-factors, or additively as a linear combination of the first phase and second phase g-factors. We use the superscript "*" to denote overall weights that is, weights taking both phases into account. The superimposed symbol "~" denotes calibrated weights.

3. *Calibration with generalized least squares distance*

The auxiliary information available at each phase of sampling can be used to obtain improved weights by the process known as calibration. The improvement translates as smaller variances of the resulting estimates. We seek a set of "new" weights that lie as close as possible to a set of starting weights. Calibration requires the specification of a distance function measuring the distance between the starting weights and the new weights. Several distance functions have been proposed, see Deville and Särndal (1992), Deville, Särndal, and Sautory (1993), and Singh (1994). Any one of these distance functions could be used for the two phase calibration procedure that we now present. We concentrate on the generalized least squares (GLS) distance defined as follows for an arbitrary set of units $k \in s$:, $\frac{1}{2}\Sigma_s \frac{c_k(\tilde{w}_k - w_k)^2}{w_k}$, where $\{w_k: k \in s\}$ are the starting weights; $\{\tilde{w}_k: k \in s\}$ are the new calibrated weights; and $\{c_k: k \in s\}$ are specified factors used to control the relative importance of the terms of the sum. We now minimize the GLS distance, successively in each phase and subject to restrictions, thereby obtaining a set of overall calibrated weights.

(I) First phase calibration (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

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

subject to

$$\Sigma_{s_1} \tilde{w}_{1k} x_{1k} = \Sigma_U x_{1k}$$

where the total $\Sigma_U x_{1k}$ is known by assumption (I). Note that this calibration cannot involve information concerning x_{2k} because it is available only up to s_1. The calibrated weights are $\tilde{w}_{1k} = w_{1k} g_{1k}$ with

$$g_{1k} = 1 + (\Sigma_U x_{1k} - \Sigma_{s_1} w_{1k} x_{1k})' T_1^{-1} x_{1k} / c_{1k} \quad (3.1)$$

for $k \in s_1$ where

$$T_1 = \Sigma_{s_1} \frac{w_{1k} x_{1k} x_{1k}'}{c_{1k}}. \quad (3.2)$$

(ii) Second phase calibration (from s_2 to s_1).

We use as starting weights $\{\tilde{w}_{1k} w_{2k}: k \in s_2\}$. This is quite reasonable because they represent a set of possible weights for making estimates from the data $\{y_k: k \in s_2\}$. Note that $\tilde{w}_{1k} w_{2k} = w_k^* g_{1k}$, where $w_k^* = w_{1k} w_{2k}$ and g_{1k} is given by (3.1). However, these starting weights do not profit from the information contained in the x_{2k}-values, available for $k \in s_2$. The second phase calibration improves the weights by incorporating this information. We consider two different formulations of the second phase calibration. They correspond to two different GLS distance functions.

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

$$\frac{1}{2}\Sigma_{s_2} \frac{c_{2k}(\tilde{w}_k^* - \tilde{w}_{1k} w_{2k})^2}{\tilde{w}_{1k} w_{2k}} \quad (3.3)$$

where $\{c_{2k}: k \in s_2\}$ are specified positive weights, subject to the second phase calibration equation

$$\Sigma_{s_2} \tilde{w}_k^* \boldsymbol{x}_k = \Sigma_{s_1} \tilde{w}_{1k} \boldsymbol{x}_k \qquad (3.4)$$

and $\boldsymbol{x}_k = (\boldsymbol{x}'_{1k}, \boldsymbol{x}'_{2k})'$. The weights resulting from this calibration define the overall calibrated weights. They are

$$\tilde{w}_k^* = \tilde{w}_{1k} w_{2k} g_{2k}^M = w_{1k}^* g_{1k} g_{2k}^M \qquad (3.5)$$

where

$$g_{2k}^M = 1 + (\Sigma_{s_1} \tilde{w}_{1k} \boldsymbol{x}_k - \Sigma_{s_2} \tilde{w}_{1k} w_{2k} \boldsymbol{x}_k)' \left(\boldsymbol{T}_2^M\right)^{-1} \boldsymbol{x}_k / c_{2k} \qquad (3.6)$$

for $k \in s_2$, and

$$\boldsymbol{T}_2^M = \Sigma_{s_2} \frac{w_k^* g_{1k} \boldsymbol{x}_k \boldsymbol{x}_k'}{c_{2k}} \qquad (3.7)$$

For $\boldsymbol{x}_{1k}$, (3.4) implies that $\Sigma_{s_2} \tilde{w}_k^* \boldsymbol{x}_{1k} = \Sigma_{s_1} \tilde{w}_{1k} \boldsymbol{x}_{1k} = \Sigma_U \boldsymbol{x}_{1k}$. In other words, for the auxiliary vector $\boldsymbol{x}_1$, the calibration carried out according to (3.4) will yield an overall weight system $\{\tilde{w}_k^*\}$ guaranteeing that the estimate of the known quantity $\Sigma_U \boldsymbol{x}_{1k}$ is exact, which is a desirable property.

For $\boldsymbol{x}_{2k}$, calibrating according to (3.4) implies that $\Sigma_{s_2} \tilde{w}_k^* \boldsymbol{x}_{2k} = \Sigma_{s_1} \tilde{w}_{1k} \boldsymbol{x}_{2k}$. Both sums are estimates of the unknown x_2-total $\Sigma_U \boldsymbol{x}_{2k}$. In other words, the calibration (3.4) assures that the first and second phase estimates of the unknown population total of $\boldsymbol{x}_2$ agree, which is also a desirable property.

As (3.5) shows, the calibration factors g_{1k} and g_{2k}^M operate multiplicatively here, resulting in the overall calibration factor $g_k^* = g_{1k} g_{2k}^M$.

One can criticize the distance function (3.3) because the factors $c_{2k} / \tilde{w}_{1k} w_{2k}$ are not necessarily all positive. Because g_{1k} can be zero or negative, the terms of (3.3) may not be all finite and positive, contradicting the notion of distance. A modified GLS distance function is therefore considered in the following Case B.

Case B (*Additive g-factors*): An alternative for the second phase calibration is to replace (3.3) by

$$\frac{1}{2} \Sigma_{s_2} \frac{c_{2k} (\tilde{w}_k^* - \tilde{w}_{1k} w_{2k})^2}{w_k^*} \qquad (3.8)$$

where $\{c_{2k} : k \in s_2\}$ are specified positive weights. Then the factors c_{2k} / w_k^* are always positive. The overall calibrated weights resulting from minimizing (3.8) subject to (3.4) is

$$\tilde{w}_k^* = w_k^* \left(g_{1k} + g_{2k}^A - 1\right) \qquad (3.9)$$

where

$$g_{2k}^A = 1 + \left(\Sigma_{s_1} \tilde{w}_{1k} \boldsymbol{x}_k - \Sigma_{s_2} \tilde{w}_{1k} w_{2k} \boldsymbol{x}_k\right)' \left(\boldsymbol{T}_2^A\right)^{-1} \boldsymbol{x}_k / c_{2k} \qquad (3.10)$$

for $k \in s_2$ with

$$\boldsymbol{T}_2^A = \Sigma_{s_2} \frac{w_k^* \boldsymbol{x}_k \boldsymbol{x}_k'}{c_{2k}} \qquad (3.11)$$

The calibration factors g_{1k} and g_{2k}^A operate additively here, and the overall calibration factor is $g_k^* = g_{1k} + g_{2k}^A - 1$.

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

$$g_k^* = \begin{cases} g_{1k} g_{2k}^M & \text{for the multiplicative case} \\ g_{1k} + g_{2k}^A - 1 & \text{for the additive case} \end{cases} \qquad (3.12)$$

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 $\boldsymbol{T}_2^A$.

Having determined the overall weights w_k^*, we use them to form the estimator of Y given by

$$\hat{Y} = \Sigma_{s_2} \tilde{w}_k^* y_k. \qquad (3.13)$$

Remark 3.1: The auxiliary data in Table 1 can be used for calibration in several ways.

Three different ways to specify the vector $\boldsymbol{x}_k$ in the second phase calibration are: (I) $\boldsymbol{x}_k = \left(\boldsymbol{x}'_{1k}, \boldsymbol{x}'_{2k}\right)'$; (ii) $\boldsymbol{x}_k = \boldsymbol{x}_{2k}$; and (iii) $\boldsymbol{x}_k = \boldsymbol{x}_{1k}$. We comment on these possibilities, assuming that the first phase calibration (3.1) is carried out in any of the three cases. The case (I) specification $\boldsymbol{x}_k = \left(\boldsymbol{x}'_{1k}, \boldsymbol{x}'_{2k}\right)'$, recommended in Särndal, Swensson, and Wretman (1992), capitalizes on all the available information. We call $\boldsymbol{x}_k = \left(\boldsymbol{x}'_{1k}, \boldsymbol{x}'_{2k}\right)'$ the *full vector*.

Both cases (ii) and (iii) disregard some available information. Case (iii) is unrealistic in that it entails observing the data $\{\boldsymbol{x}_{2k} : k \in s_1\}$, then disregarding them. We do not further consider case (iii). The case (ii) vector $\boldsymbol{x}_k = \boldsymbol{x}_{2k}$ will be called the *reduced vector*.

Second phase calibration on the reduced vector $\boldsymbol{x}_k = \boldsymbol{x}_{2k}$ can be carried out without significant loss of information if $\boldsymbol{x}_{2k}$ is a good substitute for $\boldsymbol{x}_{1k}$, as observed by Dupont (1995). By contrast, if $\boldsymbol{x}_{1k}$ complements $\boldsymbol{x}_{2k}$, then the full vector $\boldsymbol{x}_k = \left(\boldsymbol{x}'_{1k}, \boldsymbol{x}'_{2k}\right)'$ should clearly be used in (3.4). Otherwise, significant loss of information and increased variance may result.

4. *The two phase calibration estimator viewed as a regression estimator*

Alternative expressions for the calibration estimator (3.13) are given in this section. The expressions link it with the regression estimator for two phase designs introduced in Särndal Swensson and Wretman (1992, chapter 9). We show that when the weights $\tilde{w}_k^*$ are determined by either (3.5) or (3.9), then the estimator (3.13), $\hat{Y} = \sum_{s_2} \tilde{w}_k^* y_k$, can be written alternatively as

$$\hat{Y} = \sum_U \hat{y}_{1k} + \sum_{s_1} w_{1k}(\hat{y}_{2k} - \hat{y}_{1k}) + \sum_{s_2} w_k^*(y_k - \hat{y}_{2k}) \tag{4.1}$$

Here, $\hat{y}_{1k}$ and $\hat{y}_{2k}$ are successive regression predictions given as follows:

$$\hat{y}_{1k} = \boldsymbol{x}_{1k}' \hat{\boldsymbol{B}}_1 \tag{4.2}$$

with

$$\hat{\boldsymbol{B}}_1 = \boldsymbol{T}_1^{-1} \left\{ \sum_{s_1} \frac{w_{1k} \boldsymbol{x}_{1k} \hat{y}_{2k}}{c_{1k}} + \sum_{s_2} \frac{w_k^* \boldsymbol{x}_{1k}(y_k - \hat{y}_{2k})}{c_{1k}} \right\} \tag{4.3}$$

where $\mathbf{T}_1$ is given by (3.2), and

$$\hat{y}_{2k} = \boldsymbol{x}_k' \hat{\boldsymbol{B}}_2 \tag{4.4}$$

with

$$\hat{\boldsymbol{B}}_2 = \begin{cases} \left(\boldsymbol{T}_2^M\right)^{-1} \sum_{s_2} \dfrac{w_k^* g_{1k} \boldsymbol{x}_k y_k}{c_{2k}} & \text{for the multiplicative form} \\ \left(\boldsymbol{T}_2^A\right)^{-1} \sum_{s_2} \dfrac{w_k^* \boldsymbol{x}_k y_k}{c_{2k}} & \text{for the additive form} \end{cases} \tag{4.5}$$

where $\boldsymbol{T}_2^M$ and $\boldsymbol{T}_2^A$ are given by (3.7) and (3.11) respectively.

The argument showing the equivalence of (3.13) and (4.1) involves two phases of regression estimation. Suppose $\{y_k : k \in s_1\}$ were the observed y-data, and that auxiliary information on $\boldsymbol{x}_{1k}$ is available as described in assumption (I) of Section 2. Then the regression estimator of $Y = \sum_U y_k$ would be given by

$$\hat{Y} = \sum_U \hat{y}_{1k}^0 + \sum_{s_1} w_{1k}(y_k - \hat{y}_{1k}^0) \tag{4.6}$$

where $\hat{y}_{1k}^0 = \boldsymbol{x}_{1k}' \hat{\boldsymbol{B}}_1^0$, with $\hat{\boldsymbol{B}}_1^0 = \boldsymbol{T}_1^{-1} \sum_{s_1} \dfrac{w_{1k} \boldsymbol{x}_{1k} y_k}{c_{1k}}$, is the predictor of y_k based the regression of y_k on $\boldsymbol{x}_{1k}$ for $k \in s_1$. Note that $\sum_U \hat{y}_{1k}^0 = (\sum_U \boldsymbol{x}_{1k})' \hat{\boldsymbol{B}}_1^0$, where $\sum_U \boldsymbol{x}_{1k}$ is known by assumption (I). In (4.6), $\sum_{s_1} w_{1k} y_k$ represents the (hypothetical) first phase Horvitz-Thompson estimator of Y. However, neither $\sum_{s_1} w_1 y_{1k}$ nor $\hat{\boldsymbol{B}}_1^0$ can be computed because y_k is observed for the second phase sample only. A second step of regression estimation is thus necessary and is carried out as follows. In (4.6) replace the unknown $\sum_{s_1} w_{1k} y_k$ by its conditional regression estimator,

$$\sum_{s_1} w_{1k} \hat{y}_{2k} + \sum_{s_2} w_k^* (y_k - \hat{y}_{2k}) \tag{4.7}$$

where $\hat{y}_{2k} = \boldsymbol{x}_k' \hat{\boldsymbol{B}}_2$ with $\hat{\boldsymbol{B}}_2$ given by (4.5) is the predictor of y_k based on the regression of y_k on $\mathbf{x}_{2k}$, known up to s_1. Further, we replace $\hat{\boldsymbol{B}}_1^0$ in (4.6) by the regression estimator $\hat{\boldsymbol{B}}_1$ given by (4.3). With these replacements in (4.6), we obtain after some algebra the two-phase regression estimator given by (4.1).

5. *Variance Estimation*

The Taylorized variance estimator of the two-phase regression estimator $\hat{Y}$ given by (3.13), or equivalently, by (4.1) requires as a first step that we compute the residuals arising from the regression fits described in Section 4. When these residuals have been computed, the variance estimation proceeds as specified in Särndal, Swensson, and Wretman (1992), Section 9.7.

The two required sets of residuals are $e_{1k} = y_k - \boldsymbol{x}_{1k}' \hat{\boldsymbol{B}}_1$ for $k \in s_2$ and $e_{2k} = y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}_2$ for $k \in s_2$, where $\hat{\boldsymbol{B}}_1$ and $\hat{\boldsymbol{B}}_2$ are given by (4.3) and (4.5) respectively.

The variance estimator for $\hat{Y}$ is calculated as a total of two components, one for each phase, according to

$$\begin{aligned} v(\hat{Y}) = {} & \sum\sum_{s_2} w_{2k\ell} \left(w_{1k} w_{1\ell} - w_{1k\ell}\right)\left(g_{1k} e_{1k}\right)\left(g_{1\ell} e_{1\ell}\right) \\ & + \sum\sum_{s_2} w_{1k} w_{1\ell} \left(w_{2k} w_{2\ell} - w_{2k\ell}\right)\left(g_{2k} e_{2k}\right)\left(g_{2\ell} e_{2\ell}\right) \end{aligned} \tag{5.1}$$

where $w_{1k} = 1/\pi_{1k}$ and $w_{1k\ell} = 1/\pi_{1k\ell}$ with $\pi_{1k\ell} = P(k \text{ and } \ell \in s_1)$ are associated with the first phase of sampling, $w_{2k} = 1/\pi_{2k}$ and $w_{2k\ell} = 1/\pi_{2k\ell}$, with $\pi_{2k\ell} = P(k \text{ and } \ell \in s_2 | s_1)$ are their respective counterparts for the second phase. Note that for k= ℓ, we have $w_{1k\ell} = w_{1k}$, $w_{2k\ell} = w_{2k}$ in (5.1).

The g-factors g_{1k}, $g_{2k} = g_k^M$ or $g_{2k} = g_k^A$ are as defined in Section 3. Note that $\hat{\boldsymbol{B}}_2$ is as in (4.5) with two different definitions depending on whether we are in the multiplicative case or in the additive case.

6. *Domain estimation*

The overall calibrated weights $\tilde{w}_k^*$ obtained as described in Section 3 are also used to derive estimates of totals for arbitrarily specified domains. Let $U_d (U_d \subseteq U)$ be a domain of U. The y-total for the domain U_d is defined by

$Y(d) = \Sigma_{U_d} y_k = \Sigma_U y_k(d)$ with $y_k(d) = y_k$ if $k \in U_d$ and $y_k(d) = 0$ if $k \notin U_d$

Using the calibrated weights $\tilde{w}_k^*$, the estimator of $Y(d)$ is

$$\hat{Y}(d) = \Sigma_{s_2} \tilde{w}_k^* y_k(d) \qquad (6.1)$$

The variance for the domain total estimator (6.1) is obtained by the same formula (5.1), provided y_k is replaced throughout the calculation with the domain variable value $y_k(d)$. That is, e_{1k} and e_{2k} become

$$e_{1k}(d) = y_k(d) - x_{1k}' \hat{B}_1(d) \text{ for } k \in s_2$$

and

$$e_{2k}(d) = y_k(d) - x_k' \hat{B}_2(d) \text{ for } k \in s_2$$

where $\hat{B}_1(d)$ and $\hat{B}_2(d)$ are calculated from the expressions (4.3) and (4.5) for $\hat{B}_1$ and $\hat{B}_2$, replacing y_k by $y_k(d)$.

7. *Calibration groups*

We consider the case where the auxiliary data in Table 1 also includes information about membership in arbitrary subsets U_{p_1}, $p_1 = 1, \ldots, P_1$ forming a partition of the population U or subsets s_{1p_2}; $p_2 = 1, \ldots, P_2$, forming a partition of the phase one sample s_1. We designate such subsets as *calibration groups*. These calibration groups can be defined independently of one another.

Let the vector of observed auxiliary data be $x_k = (x_{1k}', x_{2k}')'$. Information for the Q_1 - dimensional vector x_{1k} is available up to the level of the full population, while information for the Q_2 - dimensional vector x_{2k}, is available up to the level of the first phase sample only. Similarly, as in Section 2, we make the following assumptions:

I) x_{1k} is known for all units $k \in U$, or $\Sigma_{U_{p_1}} x_{1k}$ is known for $p_1 = 1, \ldots, P_1$;

ii) $x_k = (x_{1k}', x_{2k}')'$ is observed for all $k \in s_1$;

iii) x_{2k} and y_k are observed for all $k \in s_2$.

The g_1-factor may be written as

$$g_{1k} = 1 + \left(\Sigma_{U_{p_1}} x_{1k} - \Sigma_{s_{1p_1}} w_{1k} x_{1k}\right)' T_{1p_1}^{-1} \frac{x_{1k}}{c_{1k}} \qquad (7.1)$$

for $k \in s_{1p_1}$, where

$$T_{1p_1} = \Sigma_{s_{1p_1}} \frac{w_k^* x_{1k} x_{1k}'}{c_{1k}}$$

Combining the additive and multiplicative definition of g_{2k} into one expression, using (3.6) and (3.10), we obtain that the g_2 -factor can be written as:

$$g_{2k} = 1 + \left(\Sigma_{s_{1p_2}} \tilde{w}_{1k} x_k - \Sigma_{s_{2p_2}} \tilde{w}_{1k} w_{2k} x_k\right)' \left(T_{2p_2}\right)^{-1} \frac{x_k}{c_{2k}} \text{ for } k \in s_{2p_2} \qquad (7.2)$$

where $T_{2p_2} = \Sigma_{s_{2p_2}} \frac{w_k^* I_k x_k x_k'}{c_{2k}}$, and

$$I_k = \begin{cases} 1 & \text{for the additive case} \\ g_{1k} & \text{for the multiplicative case} \end{cases}$$

The calibration factor g_k^* is given by (3.6) using as definitions for g_{1k} and g_{2k} expressions (7.1) and (7.2) respectively. Note that the weights $\tilde{w}_{1k} = w_{1k} g_{1k}$ and $\tilde{w}_k^* = w_k^* g_k^*$ are calibrated group by group in each of the two phases. That is

$$\Sigma_{s_{1p_1}} \tilde{w}_{1k} x_{1k} = \Sigma_{U_{p_1}} x_{1k} \text{ for } p_1 = 1, \ldots, P_1$$

and

$$\Sigma_{s_{2p_2}} \tilde{w}_k^* x_k = \Sigma_{s_{1p_2}} \tilde{w}_{1k} x_k \text{ for } p_2 = 1, \ldots, P_2$$

The resulting estimator of the total Y is

$$\hat{Y} = \Sigma_{s_2} \tilde{w}_k^* g_k^* y_k$$

The estimated variance for $\hat{Y}$ given by (5.1) requires the following residuals

$$e_{1k} = y_k - x_{1k}' \hat{B}_{1p_1} \text{ for } k \in s_{2p_1}$$

and

$$e_{2k} = y_k - x_{2k}' \hat{B}_{2p_2} \text{ for } k \in s_{2p_2}$$

where the estimated regression vectors $\hat{B}_{1p_1}$ and $\hat{B}_{2p_2}$ are

$$\hat{B}_{1p_1} = T_{1p_1}^{-1} \left\{ \Sigma_{s_{1p_1}} \frac{w_{1k} x_{1k} \hat{y}_{2k}}{c_{1k}} + \Sigma_{s_{2p_1}} \frac{w_k^* x_{1k} (y_k - \hat{y}_{2k})}{c_{1k}} \right\} \qquad (7.3)$$

and

$$\hat{B}_{2p_2} = T_{2p_2}^{-1} \Sigma_{s_{2p_2}} w_k^* x_{2k} y_k / c_{2k} \qquad (7.4)$$

Applications to domain estimation proceed as in section 6.

8. *Applications*

8.1 The case of the tax sample at Statistics Canada

One application of the approach currently in use at Statistics Canada is the two-phase design for sampling of tax records, as described in Armstrong and St-Jean (1994).

The first phase of this survey uses stratified sampling on tax records, which are then poststratified. These poststrata form the phase one calibration groups. A second phase sample is drawn from the first phase sample and re-poststratified, but by a different criterion. This second set of poststrata form the phase two calibration groups.

The estimation for this design represents an example of crossing poststratifications. Furthermore, the overall g-factors used in the survey are of the multiplicative type. The x_k-vector used in the phase two calibration procedure is the reduced one, $x_k = x_{2k}$. The first phase calibration factors g_{1k} are obtained from (7.1). We have $x_{1k} = x_{2k} = 1$ for all k.

8.2 The Case of the Canadian Survey Employment Payrolls, and Hours

The Survey on Employment Payrolls, and Hours (SEPH) covers all sectors of the Canadian industries, and collects four principal variables: (I) salaries and payments to employees (denoted as x_2; called payrolls); (ii) hours worked by employees (x_3; hours); (iii) number of employees (y_1; employment) and (iv) summarized earnings (y_2; earnings).

SEPH uses a stratified two-phase sampling design. In the first phase, a sample of payroll deduction accounts is selected using a stratified Bernoulli sampling design with sampling rates ranging from 10% to 100%. Those strata are defined by region. A region is made up of one or more Canadian provinces.

In the first phase sample, the variables x_2 and x_3 are transcribed for selected units. In the second phase, a simple random sample is drawn. Data on the two variables y_1 and y_2 are collected for respondents in this sample. In addition, classification by industry and province is recorded for sampled units, which makes it possible to derive estimates for industry-by-province domains, using the methodology described in this paper.

9. Conclusions

Our objective in this paper has been to develop the theory for two-phase sampling with regression estimation so that it can be readily incorporated into Statistics Canada's Generalized Estimation System (GES), a general purpose software that currently handles only single phase designs. We have shown how the theory is applied to estimation for arbitrary domains of the sampled finite population.

Our theory extends the results in Särndal, Swensson, and Wretman (1992), chapter 9, and in so doing, it relies on two notions important to the GES, as described in Estevao, Hidiroglou, and Särndal (1995), namely, the ideas of calibration group and regression type.

We have illustrated the theory with two specific examples at Statistics Canada, where the two phase methodology is currently used. The theory has potential application to any two phase sample design that uses auxiliary information.

Two-phase designs are powerful and economical but up until now they have been underrated. The fact that they have not been so far used much in practice has been due in part to the lack of a unified theory and appropriate software. This work should open the door for a more extensive use of two phase designs in statistical agencies.

BIBLIOGRAPHY

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

Chaudhuri, A., and Roy, D. (1994). Model Assisted Survey Sampling Strategy in Two Phases. *Metrika*, 41, 355-362.

Breidt, J. and Fuller, W.A. (1993). Regression weighting for multiplicative samples. *Sankhyá*, 55, 297-309.

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*, 87, 376-382.

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

Dupont, F. (1995). Redressement alternatifs en présence de plusieurs niveaux d'information auxillaire. Internal report from INSEE, Paris, France.

Estevao, V., Hidiroglou, M.A., and Särndal, C.E. (1995). Requirements on a Generalized Estimation System at Statistics Canada. *Journal of Official Statistics*.

Hidiroglou, M.A., Latouche, M., Armstrong, B., and Gossen, M. (1995). Improving survey Information Using Administrative Records: The Case of the Canadian Employment Survey. Paper presented at the Annual research Conference, held in Washington.

Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika*, 6, 125-133.

Särndal, C.E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

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

Singh, A. (1994). Sampling design-based estimating functions for finite population parameters. Annual Meeting of the Statistical Society of Canada, Banff, Alberta, May 8-11.