Survey Methodology, June 1998
Vol. 24, No. 1, pp. 11-20
Statistics Canada

Use of Auxiliary Information for Two-phase Sampling

M.A. HIDIROGLOU and C.-E. SÄRNDAL[1]

ABSTRACT

Two-phase sampling designs offer a variety of possibilities for use of auxiliary information. We begin by reviewing the different forms that auxiliary information may take in two-phase surveys. We then set up the procedure by which this information is transformed into calibrated weights, which we use to construct efficient estimators of a population total. The calibration is done in two steps: (i) at the population level; (ii) at the level of the first-phase sample. We go on to show that the resulting calibration estimators are also derivable via regression fitting in two steps. We examine these estimators for a special case of interest, namely, when auxiliary information is available for population subgroups called calibration groups. Poststrata are the simplest example of such groups. Estimation for domains of interest and variance estimation are also discussed. These results are illustrated by applying them to two important two-phase designs at Statistics Canada. The general theory for using auxiliary information in two-phase sampling is being incorporated into Statistics Canada's Generalized Estimation System.

KEY WORDS: Generalized regression; Two-phase sampling; Model assisted approach; Domain estimation; Calibration factors.

1. INTRODUCTION

Two-phase sampling is a powerful and cost-effective technique. It was first proposed by Neyman (1938). In Cochran's (1977) book, and in its two earlier editions dated 1953 and 1963, one finds basic results for two-phase sampling, including the simplest regression estimators for such designs. This paper takes a broader outlook and proposes a general approach to the use of auxiliary information in two-phase survey designs. Our main references are Särndal and Swensson (1987), Särndal, Swensson and Wretman (1992) and Dupont (1995). Recent related work includes Breidt and Fuller (1993), who presented computationally efficient estimation procedures for three-phase sampling in the presence of auxiliary information. Chaudhuri and Roy (1994) studied optimality properties of the well-known simpler regression estimators for two-phase sampling. Binder (1996) described a simple linearization procedure to estimate variances of nonlinear estimators. His procedure can be applied to any sampling design, including two-phase-sampling. Throughout this paper, we assume *arbitrary* sampling designs for each of the two phases.

Single-phase sampling involves the use of one layer of information for estimation. In two-phase sampling, however, one has to consider two layers of information. This complicates matters, and it is not clear-cut how best to exploit the combined information from the two sources. Two approaches are considered in this paper for building estimators based on auxiliary information. These are the *calibration approach* and the *generalized regression approach.* We show that the generalized regression approach can be viewed as a special case of the calibration approach. The two approaches are examined under a common structure for the auxiliary information. It assumes that information exists about an auxiliary vector x_1 for the units of the entire population, and about a second auxiliary vector x_2 for the units of the first phase sample. Consequently, at the level of the first phase sample, there is information about both vectors, x_1 and x_2.

The *generalized regression approach,* as applied to two-phase sampling, is discussed in Särndal *et al.* (1992). These authors develop the general regression estimator for two-phase sampling, assuming arbitrary sampling designs in each of the two phases. Two regression fits are carried out. A "bottom level" regression is fitted to produce predicted values up to the level of the first phase sample, using the auxiliary information available for this step. Next, a "top level" regression is fitted to produce predicted values up to the entire population level, using the information appropriate for this step. The two sets of predicted values are used to build a generalized regression estimator.

The *calibration approach* focuses on the weights given to the units for purposes of estimation. Calibration implies that a set of starting weights (usually the sampling design weights) are transformed into a set of new weights, called calibrated weights. The calibrated weight of a unit is the product of its initial weight and a calibration factor. The calibration factors are obtained by minimizing a function measuring the distance between the initial weights and the calibrated weights, subject to the constraint that the calibrated weights yield exact estimates of the known auxiliary population totals. In two-phase sampling the two levels of information imply two consecutive calibrations. The first phase of calibration uses the auxiliary information available (at least population counts) at the level of the entire

[1] M.A. Hidiroglou, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6; and C.-E. Särndal, University of Montreal, and Statistics Canada.

population, resulting in first-phase calibrated weights. The second phase of calibration uses these first-phase calibrated weights and incorporates the information at the level of the first-phase sample, resulting in a final set of calibrated weights.

Both approaches profit from the two layers of information. They do not necessarily yield identical results. Whether they do or not depends on the exact formulations given to the regression fits and the calibration approach. This is apparent in Dupont (1995), where four alternative estimators were developed under the regression approach. These differ in the way that the auxiliary variables are used in deriving the predicted y-values required for the regression estimator. For each of these four approaches, Dupont built a matching estimator using the calibration approach. She succeeded in obtaining an exact equivalence between the two approaches in only one of the four cases. Three of Dupont's four approaches can be considered as special cases of the general approach in this paper.

In this paper, building on Hidiroglou and Särndal (1995), we provide a unified theory for two-phase sampling with auxiliary information. We show that the regression estimators can be obtained as a special case of the calibration approach. Direct linkage between the two approaches is therefore possible. One motivation for our work was the necessity to provide tools for efficient use of administrative data sources in several important Statistics Canada surveys. Our work has prepared the way for the inclusion of two-phase sampling into Statistics Canada's Generalized Estimation System described in Estevao, Hidiroglou and Särndal (1995).

We illustrate our general theory by applying it to two survey designs currently used at Statistics Canada. The first application, Armstrong and St-Jean (1994), describes the use of the two-phase approach for sampling tax records. Our second application, Hidiroglou, Latouche, Armstrong and Gossen (1995), involves the use of two-phase sampling of payroll deduction accounts used in Statistics Canada's Survey of Earnings, Payrolls and Hours.

The paper is organized as follows. Section 2 sets up the notation. Section 3 specifies our version of the calibration approach in two-phase sampling. Section 4 establishes the important result that the resulting calibration estimator can be expressed, with exact equivalence, as a two-phase regression estimator, that is, one derived via two consecutive regression fits. Additional theoretical results are reported in Sections 5 and 6. Section 5 examines the forms taken by our two-phase calibration estimator under important special types of information, namely, when some of the auxiliary variables, either in the first or in the second phase, correspond to categorical variables that codify a grouping of the units into mutually exclusive and exhaustive classes. Section 6 gives results on two issues that always require attention in a survey, which are central to the GES, namely, (a) estimation for domains (sub-populations), and (b) design-based variance estimation. For variance estimation we use the approach of Särndal and Swensson (1987). Section 7 shows how the preceding theory is applied to two-phase designs currently in use at Statistics Canada. Finally, Section 8 provides a brief summary.

2. 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, according to a sampling design with the selection probabilities $\pi_{1k} = P(k \in s_1)$. Given s_1, a second-phase sample $s_2\,(s_2 \subseteq s_1 \subseteq U)$ is selected from s_1, according to a sampling design with the selection probabilities $\pi_{2k} = P(k \in s_2 \mid s_1)$. Note that these are conditional probabilities, given s_1. We assume that $\pi_{1k} > 0$ for all $k \in U$ and $\pi_{2k} > 0$ for all $k \in s_1$. From this point on, we work with weights in the estimation process. We will denote the first-phase sampling weight of unit k as $w_{1k} = 1/\pi_{1k}$, and the second-phase sampling weight as $w_{2k} = 1/\pi_{2k}$. The overall sampling weight for a selected unit is $w_k^* = w_{1k}w_{2k}$.

Our objective is to estimate the population total $Y = \sum_U y_k$, where y_k is the value of the variable of interest y for unit k. If $A \subseteq U$ is an arbitrary set of units, we write simply $\sum_A$ for $\sum_{k \in A}$. The customary two-phase sampling procedure calls for collecting inexpensive information about the units k belonging to a large first-phase sample s_1. This information is then used to realize efficient sampling and estimation in the second phase. The values y_k are recorded for $k \in s_2$. An unbiased estimator of Y is given by $\hat{Y} = \sum_{s_2} w_k^* y_k$. This estimator uses sampling weights only. A more extensive use of available auxiliary information is achieved through the regression estimators that we will now examine.

We denote the auxiliary vector at the level of the first-phase sample as $\boldsymbol{x}$ and its value for unit k as $\boldsymbol{x}_k$. As in Särndal *et al.* (1992, chapter 9), we partition $\boldsymbol{x}_k$ as $\boldsymbol{x}_k = (\boldsymbol{x}_{1k}', \boldsymbol{x}_{2k}')'$. Information is available up to the entire population level for the vector $\boldsymbol{x}_{1k}$, whereas for the vector $\boldsymbol{x}_{2k}$, information is only available up to the level of the first-phase sample. Table 1 summarizes our assumptions on the auxiliary information available for estimation.

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

Set of units	Data available
Population	$\{\boldsymbol{x}_{1k} : k\, \varepsilon\, U\}$ or $\sum_U \boldsymbol{x}_{1k}$
First-phase sample	$\{\boldsymbol{x}_k : k\, \varepsilon\, s_1\}$
Second-phase sample	$\{(\boldsymbol{x}_k, y_k) : k\, \varepsilon\, s_2\}$

Note that individual values $\boldsymbol{x}_{1k}$, $k\, \varepsilon\, U$, are not required. It suffices to know the total $\sum_U \boldsymbol{x}_{1k}$, which may be taken from a reliable administrative source. The presence of auxiliary information in one or both phases opens the possibility of modifying the sampling weights with the aid of calibration factors calculated using the auxiliary information. In each of the two phases, a unit's sampling weight is modified by multiplying it by the calibration factor, resulting in a calibrated weight.

The first-phase calibrated weight $\tilde{w}_{1k}$ is computed for units $k \in s_1$ as $\tilde{w}_{1k} = w_{1k} g_{1k}$. The first-phase sampling weight is w_{1k}, and the first-phase calibration factor is g_{1k}. Similarly, we compute overall calibration weights $\tilde{w}_k^* = w_k^* g_k^*$ for units $k \in s_2$, where g_k^* is the overall calibration factor. The superscript "*" denotes overall weights taking into account both phases. The superimposed symbol "~" denotes calibrated weights.

3. CALIBRATION WITH GENERALIZED LEAST SQUARES DISTANCE

Auxiliary information available at each phase of sampling can improve weights by the process known as calibration. This improvement yields smaller variances of the resulting estimates if there is a strong correlation between the auxiliary variables and the variables of interest. We seek a set of "new" weights that lie as close as possible to a set of starting weights. The calibration requires the specification of a measure of 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 and Mohl (1996). Any one of these distance functions could be used for two-phase calibration. However, we concentrate on one of these, namely, the generalized least squares (GLS). For an arbitrary set of units s, it is of the form

$$D = \frac{1}{2} \sum_s C_k \frac{(\tilde{w}_k - w_k)^2}{w_k} \tag{3.1}$$

where $\{w_k : k \varepsilon s\}$ are the starting weights, $\{\tilde{w}_k : k \varepsilon s\}$ are the new calibrated weights, and $\{C_k : k \varepsilon s\}$ are specified positive factors that control the relative importance of the terms of this sum. For each of the two phases, we minimize a GLS distance measure with suitable factors C_k, subject to constraints. After applying the two successive calibrations, we have a set of overall calibrated weights.

(i) **First-phase calibration** (from s_1 to U).

The first-phase sampling weights $\{w_{1k} : k \varepsilon s_1\}$ are used as starting weights. Let $\{C_{1k} : k \varepsilon s_1\}$ be pre-specified positive factors. We determine the first-phase calibrated weights by minimizing the GLS distance

$$D_1 = \frac{1}{2} \sum_{s_1} C_{1k} \frac{(\tilde{w}_{1k} - w_{1k})^2}{w_{1k}} \tag{3.2}$$

subject to the first-phase calibration equation

$$\sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_{1k} = \sum_U \boldsymbol{x}_{1k} \tag{3.3}$$

where the total $\sum_U \boldsymbol{x}_{1k}$ is known. Note that this calibration does not involve information concerning $\boldsymbol{x}_{2k}$ because it is available only up to s_1.

The resulting weights are

$$\tilde{w}_{1k} = w_{1k} g_{1k} \tag{3.4}$$

with

$$g_{1k} = 1 + \left(\sum_U \boldsymbol{x}_{1k} - \sum_{s_1} w_{1k} \boldsymbol{x}_{1k}\right)' \boldsymbol{T}_1^{-1} \frac{\boldsymbol{x}_{1k}}{C_{1k}} \tag{3.5}$$

and

$$\boldsymbol{T}_1 = \sum_{s_1} \frac{w_{1k} \boldsymbol{x}_{1k} \boldsymbol{x}_{1k}'}{C_{1k}} \tag{3.6}$$

Some of the $\tilde{w}_{1k}$ given by (3.4) may be negative, or zero. Many users prefer weights to be always positive. This can be achieved by adding to (3.3) the inequality constraints $\tilde{w}_{1k} > 0$ for all $k \in s_1$. The resulting weights have no closed expression, in contrast to (3.4).

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

We use $\{\tilde{w}_{1k} w_{2k}, k \varepsilon s_2\}$ as starting weights, where $\tilde{w}_{1k}$ is given by (3.4). These weights incorporate the information about $\boldsymbol{x}_{1k}$ available up to the full population level. Applying them to the data $\{y_k : k \varepsilon s_2\}$ yields one possible estimator, namely $\hat{Y} = \sum_{s_2} \tilde{w}_{1k} w_{2k} y_k$. However, since these weights do not contain the $\boldsymbol{x}_{2k}$-value information available for $k \varepsilon s_1$, they can be improved through a second-phase calibration. Let $\{C_{2k} : k \varepsilon s_2\}$ be specified positive factors. We determine the overall calibrated weights $\tilde{w}_k^*$ by minimizing

$$D_2 = \frac{1}{2} \sum_{s_2} \frac{C_{2k}(\tilde{w}_k^* - \tilde{w}_{1k} w_{2k})^2}{\tilde{w}_{1k} w_{2k}} \tag{3.7}$$

subject to the second-phase calibration equation

$$\sum_{s_2} \tilde{w}_k^* \boldsymbol{x}_k = \sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_k \tag{3.8}$$

where $\boldsymbol{x}_k = (\boldsymbol{x}_{1k}', \boldsymbol{x}_{2k}')'$. The resulting overall calibrated weights are

$$\tilde{w}_k^* = w_k^* g_k^* \tag{3.9}$$

where

$$g_k^* = g_{1k} g_{2k} \tag{3.10}$$

with g_{1k} given by (3.5) and g_{2k} by

$$g_{2k} = 1 + \left(\sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_k - \sum_{s_2} \tilde{w}_{1k} w_{2k} \boldsymbol{x}_k\right)' \boldsymbol{T}_2^{-1} \frac{\boldsymbol{x}_k}{C_{2k}} \tag{3.11}$$

for $k \varepsilon s_2$, and

$$T_2 = \sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} \mathbf{x}_k \mathbf{x}'_k}{C_{2k}} \tag{3.12}$$

Again, some g_k^* may be zero or negative, but always positive g_k^* can be ascertained by adding to (3.8) the inequality constraints $w_k^* > 0$ for $k \varepsilon s_2$.

Having determined the overall weights $\tilde{w}_k^*$ by equation (3.9), the estimator of Y is given by

$$\hat{Y} = \sum_{s_2} \tilde{w}_k^* y_k \tag{3.13}$$

Remark 3.1 A potential problem with the above approach is that some of the g_{1k}'s may be negative or even zero. If this occurs, (3.7) is not a proper distance measure. Some of the important applications, such as poststatification, do not have this problem as their associated g_{1k}'s are always greater than zero. If all the g_{1k}'s are greater than zero, then the minimization criterion given by (3.7) is acceptable. Otherwise, we have to modify it. One possible modification is to impose on the above-mentioned constraints that the w_{1k}'s are positive for $k \varepsilon s_1$. Another possible modification is to replace C_{2k} in (3.7) by

$$C^*_{2k} = C_{2k} \frac{\tilde{w}_{1k}}{w_{1k}}.$$

Then

$$\frac{C^*_{2k}}{\tilde{w}_{1k} w_{2k}} = \frac{C_{2k}}{w_k^*},$$

which is always positive. The resulting g_k^*-factors in (3.9) can be shown to be $g_k^* = g_{1k} + g_{2k} - 1$, where g_{1k} is given as before by (3.5), and g_{2k} by (3.11) provided that we instead define $\mathbf{T}_2$ as

$$\mathbf{T}_2 = \sum_{s_2} \frac{w_k^* \mathbf{x}_k \mathbf{x}'_k}{C_{2k}}.$$

It is our opinion that in most applications the choice between the multiplicative $g_k^* = g_{1k} g_{2k}$ and the additive form $g_k^* = g_{1k} + g_{2k} - 1$ would have little effect on the resulting estimates. That is, we believe the two point estimates would be very close, and so would be their associated estimates of variance.

Remark 3.2: Bounding the weights ordinarily has negligible impact on the estimates. Recent experience with calibration for single phase designs, Stukel, Hidiroglou, and Särndal (1996), has shown that mildly different sets of g-weights lead to point estimates that differ very little. Some recently developed computer software for calibration, for example, the software described in Deville *et al.* (1993), minimizes a distance function such that the resulting g-factors are guaranteed to be bounded from above and from below.

Remark 3.3: The auxiliary data in Table 1 can be used in several ways for two-phase calibration. Considering in particular the second-phase calibration equation defined by (3.8), three different specifications of the vector $\mathbf{x}_k$ are: (i) $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$; (ii) $\mathbf{x}_k = \mathbf{x}_{2k}$; and (iii) $\mathbf{x}_k = \mathbf{x}_{1k}$. We comment on these possibilities, assuming for each of these that a first-phase calibration has been carried out, resulting in the first-phase calibrated weights (3.4).

The case (i) specification $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$, recommended in Särndal *et al.* (1992), capitalizes on all the available information. Thus, in this respect case (i) is ideal. Cases (ii) and (iii) disregard some available information. Case (ii) is sometimes of interest, despite some loss of information; an example is given in Section 7.1. Case (iii) implies that the data $\{\mathbf{x}_{2k} : k \varepsilon s_1\}$ are observed, but not used: we do not further consider this case. We call $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$ the *full vector* and $\mathbf{x}_k = \mathbf{x}_{2k}$ the *reduced vector*.

Second-phase calibration on the reduced vector $\mathbf{x}_k = \mathbf{x}_{2k}$ can be carried out without significant loss of information if $\mathbf{x}_{2k}$ is a good *substitute* for $\mathbf{x}_{1k}$, as also observed by Dupont (1995). However, if x_{1k} *complements* x_{2k}, then the full vector $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$ should clearly be used in the calibration defined by (3.7). Otherwise, significant loss of information and increased variance may result.

Remark 3.4: Both the full and the reduced $\mathbf{x}_k$-vectors lead to overall weights $\tilde{w}_k^*$ calibrated on $\mathbf{x}_{2k}$ from s_2 to s_1. This means that $\sum_{s_2} \tilde{w}_k^* \mathbf{x}_{2k} = \sum_{s_1} \tilde{w}_{1k} \mathbf{x}_{2k}$, because (3.8) holds, and $\mathbf{x}_{2k}$ is contained in $\mathbf{x}_k$. However, there exists a difference between the full and reduced vector specifications with respect to the calibration on $\mathbf{x}_{1k}$. If the full vector specification is used in phase two, the resulting overall weights $\tilde{w}_k^*$ are calibrated on $\mathbf{x}_{1k}$ from s_2 to s_1, and from s_1 to U. This means that $\sum_{s_2} \tilde{w}_k^* \mathbf{x}_{1k} = \sum_{s_1} \tilde{w}_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. In contrast, if the reduced vector specification is used, the resulting overall weights $\tilde{w}_k^*$ **are** calibrated on $\mathbf{x}_{1k}$ from s_1 to U by virtue of the first-phase calibration. That is $\sum_{s_1} \tilde{w}_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. However, they are **not** calibrated from s_2 to s_1, because $\mathbf{x}_{1k}$ is not present in the second-phase calibration. Hence, $\sum_{s_2} \tilde{w}_k^* \mathbf{x}_{1k} \neq \sum_{s_1} \tilde{w}_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. Thus if the survey requires a weight system that will reproduce the known $\sum_U \mathbf{x}_{1k}$, then the full vector specification must be used.

So far, we have focused on the general framework for calibration with two levels of auxiliary information. This framework does not reveal the many interesting forms that the estimator $\hat{Y}$ given by (3.13) may take for specific cases of auxiliary information. Some illustrations are given in Section 7. We first address three issues that are of practical interest in virtually every major survey: (i) poststratification or, more generally, the presence of auxiliary information for population subgroups (Section 5), (ii) estimation for domains of interest (Section 6), and (iii) the construction of variance estimates (Section 6).

4. THE TWO-PHASE CALIBRATION ESTIMATOR VIEWED AS A REGRESSION ESTIMATOR

An alternative expression for the calibration estimator (3.13) is given by formula (4.1) below. This expression links it exactly with the regression estimator for two-phase designs introduced in Särndal *et al.* (1992, chapter 9).

Theorem 4.1: When the overall calibrated weights $\tilde{w}_k^*$ are determined by (3.9), the calibration estimator (3.13) is identical to the two-phase regression estimator given by

$$\hat{Y} = \sum\nolimits_U \hat{y}_{1k} + \sum\nolimits_{s_1} w_{1k}(\hat{y}_{2k} - \hat{y}_{1k}) + \sum\nolimits_{s_2} w_k^*(y_k - \hat{y}_{2k}) \quad (4.1)$$

where $\hat{y}_{1k}$ and $\hat{y}_{2k}$ are successive regression predictions such that

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

with

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

where $\boldsymbol{T}_1$ is given by (3.6), and

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

with

$$\hat{\boldsymbol{B}}_2 = \boldsymbol{T}_2^{-1} \sum\nolimits_{s_2} \frac{\tilde{w}_{1k} w_{2k} \boldsymbol{x}_k y_k}{C_{2k}} \qquad (4.5)$$

where $\boldsymbol{T}_2$ is given by (3.12).

The proof for Theorem 4.1 uses some tedious but straightforward algebra and is not presented here.

We now show that (4.1) can be constructed via regression estimation in two steps. For the first step, suppose that the variable of interest y_k were observed for the full first-phase sample s_1. The auxiliary information on $\boldsymbol{x}_{1k}$ is available for $k \in s_1$ and the population total $\sum_U \boldsymbol{x}_{1k}$ is known. The resulting regression estimator of $Y = \sum_U y_k$ would then be given by

$$\hat{Y} = \sum\nolimits_U \hat{y}_{1k}^0 + \sum\nolimits_{s_1} w_{1k}\left(y_k - \hat{y}_{1k}^0\right)$$

$$= \sum\nolimits_{s_1} w_{1k}\, y_k + \left(\sum\nolimits_U \hat{y}_{1k}^0 - \sum\nolimits_{s_1} w_{1k}\, \hat{y}_{1k}^0\right) \quad (4.6)$$

In the last expression, the first term represents the (hypothetical) first-phase Horvitz-Thompson estimator of Y. The second and third terms represent a regression adjustment, where $\hat{y}_{1k}^0$ is the predictor of y_k based on the fitted regression of y_k on $\boldsymbol{x}_{1k}$ for $k \varepsilon s_1$. That is, $\hat{y}_{1k}^0 = \boldsymbol{x}_{1k}' \hat{\boldsymbol{B}}_1^0$, with

$$\hat{\boldsymbol{B}}_1^0 = \boldsymbol{T}_1^{-1} \sum\nolimits_{s_1} \frac{w_{1k}\, \boldsymbol{x}_{1k}\, y_k}{C_{1k}}.$$

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. However, none of the terms in (4.6) can be computed directly, because y_k is only observed for the second-phase sample. A second step of regression estimation is thus necessary. It is carried out by replacing the unknown $\sum_{s_1} w_{1k} y_k$ in (4.6) by its conditional regression estimator

$$\sum\nolimits_{s_1} w_{1k} \hat{y}_{2k} + \sum\nolimits_{s_2} w_k^* \, (y_k - \hat{y}_{2k}) \qquad (4.7)$$

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 $\boldsymbol{x}_k$, known up to s_1. Next, the vector $\hat{\boldsymbol{B}}_1^0$ required for computing $\hat{y}_{1k}^0$ contains a known matrix $\boldsymbol{T}_1$ and an unknown vector

$$\sum\nolimits_{s_1} \frac{w_{1k}\, \boldsymbol{x}_{1k} y_k}{C_{1k}}.$$

Using a regression estimator for this unknown vector, we obtain $\hat{\boldsymbol{B}}_1$ given by (4.3) as a replacement for $\hat{\boldsymbol{B}}_1^0$. These two substitutions in (4.6) lead to the two-phase regression estimator given by (4.1), which is identical to the calibration estimator (3.13).

Remark 4.1: A more direct alternative to $\hat{\boldsymbol{B}}_1$ in (4.3) would be to use only the second-phase sample. This would have produced

$$\hat{\boldsymbol{B}}_{1,\text{alt}} = \left(\sum\nolimits_{s_2} \frac{w_k^* \boldsymbol{x}_{1k} \boldsymbol{x}_{1k}'}{C_{2k}} \right)^{-1} \sum\nolimits_{s_2} \frac{w_k^*\, \boldsymbol{x}_{1k} y_k}{C_{2k}}$$

The resulting predictions $\hat{y}_{1k,\text{alt}} = \boldsymbol{x}_{1k}' \hat{\boldsymbol{B}}_{1,\text{alt}}$ would be replacing $\hat{y}_{1k}$ in (4.1). However, the resulting regression estimator is not identical to (3.13) and is a less efficient alternative, because $\hat{\boldsymbol{B}}_{1,\text{alt}}$ uses less $\boldsymbol{x}_{1k}$-information than $\hat{\boldsymbol{B}}_1$.

5. CALIBRATION GROUPS

In this Section we apply the results of Sections 3 and 4 to the important case where the auxiliary data in Table 1 include information about mutually exclusive and exhaustive subsets of the population U, and of the first-phase sample s_1. The population subsets are denoted by $U_i, i = 1, ..., I$, and the first-phase subsets by $s_{1j}, j = 1, ..., J$. Such subsets are called calibration groups, for reasons that will become clear later in this Section. Simple examples of calibration groups are poststrata.

Two vectors denoted Δ_{1k} and Δ_{2k} will be used to specify the membership of a given unit k in the calibration groups U_i and s_{1j}, respectively. These group identifiers are

$$\Delta_{1k} = (\delta_{11k}, ..., \delta_{1ik}, ..., \delta_{1Ik})' \qquad (5.1)$$

with

$$\delta_{1ik} = \begin{cases} 1 & \text{if } k \varepsilon U_i \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, ..., I \qquad (5.2)$$

and

$$\Delta_{2k} = (\delta_{21k}, ..., \delta_{2jk}, ..., \delta_{2Jk})' \qquad (5.3)$$

with

$$\delta_{2jk} = \begin{cases} 1 & \text{if } k \varepsilon s_{1j} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } j = 1, ..., J \qquad (5.4)$$

Besides the group membership information, which is qualitative and specified by Δ_{1k} and Δ_{2k}, there may exist information for the unit k about quantitative (continuous or discrete) variables. We call them *supplementary auxiliary variables*. For example, categorical information about a unit (enterprise) in a business survey may consist of an industry code or a geographical location code. In addition, quantitative variable information may also be available concerning the number of employees or the gross business income of the unit. Some of these supplementary auxiliary variables may be known up to the level of the population, and others up to the level of the first-phase sample.

We assume in this Section that the vector x_{1k}, used in calculating the first-phase g-factors, has the structure

$$x'_{1k} = \Delta'_{1k} \otimes z'_{1k} \qquad (5.5)$$

where z_{1k} of dimension Q_1 is the vector of supplementary auxiliary variables available for the first-phase sample. The information requirements in Table 1 apply to the vector x_{1k}. This implies that we must know either the group membership specified by Δ_{1k} and the value of z_{1k} for every $k \in U$, or the total $\sum_{U_i} z_{1k}$ separately for each group, $i = 1, ..., I$.

When x_{1k} has the form given by (5.5), the first-phase g-factors g_{1k} in (3.5) can be obtained by a group by group calculation. The T_1 matrix to be inverted, given by (3.6), is block diagonal and of dimension IQ_1 by IQ_1. The typical diagonal block, denoted as T_{1i} of dimension Q_1 by Q_1, is given by

$$T_{1i} = \sum_{s_{1i}} \frac{w_{1k}\, z_{1k} z'_{1k}}{c_{1k}} \qquad (5.6)$$

for $i = 1, ..., I$. The resulting inverse of T_1 is also block diagonal with diagonal matrices T_{1i}^{-1}. The off diagonal blocks of the inverse of T_1 are zero matrices. So we obtain from (3.6)

$$g_{1k} = 1 + \left(\sum_{U_i} z_{1k} - \sum_{s_{1i}} w_{1k} z_{1k}\right)' T_{1i}^{-1} \frac{z_{1k}}{c_{1k}} \qquad (5.7)$$

for $k \in s_{1i}$, $i = 1, ..., I$, where T_{1i} is given by (5.6). Note that the resulting weights $\tilde{w}_{1k}$ are the same as those obtained by carrying out the first-phase calibration group by group, calibrating for group i on the known total $\sum_{U_i} z_{1k}$. That is, $\sum_{s_{1i}} \tilde{w}_{1k} z_{1k} = \sum_{U_i} z_{1k}$ for $i = 1, ..., I$. It is thus fitting to call the groups U_i *first-phase calibration groups*.

Now consider the second-phase g-factors g_{2k} given by (3.11). They are based on the auxiliary vectors x_k, required to be known for the units $k \in s_1$. We assume that x_k contains information about the second-phase groups so that

$$x'_k = \Delta'_{2k} \otimes z'_k \qquad (5.8)$$

where Δ_{2k} is the second-phase group identifier, and z_k is the value of a vector of supplementary auxiliary variables available for $k \in s_1$. Since the requirements in Table 1 apply, it follows that Δ_{2k} (the second-phase group membership) and the value of z_k (the supplementary auxiliary vector) must be known for every $k \in s_1$. Here z_k may contain some or all of the information in x_{1k} given by (5.5), and any other information available for the units $k \in s_1$.

When x_k has the structure (5.8), the factors g_{2k} can also be obtained through a group by group calculation. This simplification is a result of the fact that the matrix to be inverted in (3.11) is block diagonal. We obtain

$$g_{2k} = 1 + \left(\sum_{s_{1j}} \tilde{w}_{1k} z_k - \sum_{s_{2j}} \tilde{w}_{1k} w_{2k} z_k\right)' T_{2j}^{-1} \frac{z_k}{c_{2k}} \qquad (5.9)$$

for $k \in s_{2j} = s_2 \cap s_{1j}$, $j = 1, ..., J$, where

$$T_{2j} = \sum_{s_{2j}} \frac{\tilde{w}_{1k} w_{2k} z_k z'_k}{c_{2k}} \qquad (5.10)$$

The resulting overall weights $\tilde{w}_k^* = w_k^* g_k^*$ where $g_k^* = g_{1k} g_{2k}$ are the same as those obtained by carrying out the second-phase calibration group by group, calibrating for group j on the known quantity $\sum_{s_{1j}} \tilde{w}_{1k} z_k$. That is, $\sum_{s_{2j}} \tilde{w}_k^* z_k = \sum_{s_{1j}} \tilde{w}_{1k} z_k$ for $j = 1, ..., J$. The groups s_{1j} are called *second-phase calibration groups*. We now have a procedure for computing g_{1k} and g_{2k} group by group using (5.7) and (5.9). The total Y is still estimated according to (3.13).

6. DOMAIN ESTIMATION AND VARIANCE ESTIMATION

The preceding sections dealt with estimation of the total of y at the entire population level. In most surveys, there is also a need to provide estimates for various subpopulations or domains of interest. Requests for domain estimates can be made either before or after the sampling stage of the survey. Auxiliary information is essential for domains. A precise domain estimate may be obtained (even for small domains) if: (i) calibration groups and domains of interest agree closely, and (ii) the auxiliary variables exhibit a strong regression relationship with the variable(s) of interest.

Denote by $U_d\,(U_d \subseteq U)$ any domain of the population U for which an estimate is required. The y-total for the domain U_d is defined by $Y(d) = \sum_{U_d} y_k = \sum_U y_k(d)$ with $y_k(d) = y_k$ if $k \varepsilon U_d$ and $y_k(d) = 0$ if $k \notin U_d$.

The estimator of $Y(d)$ is

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

where the overall calibrated weights $\tilde{w}_k^* = w_k^* \, g_k^*$ may be calculated group by group as described in Section 5. The calibration factors g_{1k} and g_{2k} are calculated using all relevant available auxiliary information, specified as in Table 1. So in this sense, the resulting overall calibrated weights $\tilde{w}_k^*$ are the best possible ones. Note that these weights are independent of the particular domains requiring estimation in the survey.

The estimator of the variance for the domain total estimator $\hat{Y}(d)$ is obtained using a design-based approach. This means that the variance is interpreted with reference to repeated draws of samples s_1 and s_2. Details for the derivation of this variance are given in Särndal *et al.* (1992) (Result 9.7.1, p. 362). The first order and second order inclusion probabilities enter into the weights used in the variance formula. The weights associated with the first-phase sample are $w_{1k} = 1/\pi_{1k}$ and $w_{1k\ell} = 1/\pi_{1k\ell}$ with $\pi_{1k\ell} = P(k \text{ and } \ell \in s_1)$. The weights $w_{2k} = 1/\pi_{2k}$ and $w_{2k\ell} = 1/\pi_{2k\ell}$ with $\pi_{2k\ell} = P(k \text{ and } \ell \in s_2 \mid s_1)$ denote their second phase counterparts. Two sets of regression residuals, one for each phase, are also required. The estimator of the variance of $\hat{Y}(d)$ is given by

$$v\{\hat{Y}(d)\} =$$

$$\sum_{k \varepsilon s_2} \sum_{\ell \varepsilon s_2} w_{2k\ell}(w_{1k}w_{1\ell} - w_{1k\ell})(g_{1k}e_{1k}(d))\,(g_{1\ell}e_{1\ell}(d)) + \qquad (6.2)$$

$$\sum_{k \varepsilon s_2} \sum_{\ell \varepsilon s_2} w_{1k}w_{1\ell}(w_{2k}w_{2\ell} - w_{2k\ell})\,(g_{2k}e_{2k}(d))(g_{2\ell}e_{2\ell}(d))$$

Note that for $k = \ell$ we have $w_{1k\ell} = w_{1k}$, and $w_{2k\ell} = w_{2k}$ in (6.2). We now specify the regression residuals in (6.2) assuming that there are first-phase calibration groups $U_i, i = 1, \ldots, I,$ and second-phase calibration groups $s_{1j}, j = 1, \ldots, J,$ as explained in Section 5. We denote the associated sample subsets as follows: $s_{2i} = s_2 \cap U_i$; $s_{2j} = s_2 \cap s_{1j}$. The required residuals in (6.2) are, for $k \in (s_{2i} \cap U_d)$,

$$e_{1k}(d) = y_k(d) - z_{1k}' \hat{\boldsymbol{B}}_{1i}(d) \qquad (6.3)$$

and, for $k \in (s_{2j} \cap U_d)$

$$e_{2k}(d) = y_k(d) - z'_k \hat{\boldsymbol{B}}_{2j}(d) \qquad (6.4)$$

The estimated regression vectors $\hat{\boldsymbol{B}}_{1i}(d)$ and $\hat{\boldsymbol{B}}_{2j}(d)$ are

$$\hat{\boldsymbol{B}}_{1i}(d) = \boldsymbol{T}_{1i}^{-1}$$

$$\left\{ \sum\nolimits_{s_{1i}} \frac{w_{1k} z_{1k} \hat{y}_{2k}(d)}{C_{1k}} + \sum\nolimits_{s_{2i}} \frac{w_k^* \, z_{1k}(y_k(d) - \hat{y}_{2k}(d))}{C_{1k}} \right\} (6.5)$$

where $\boldsymbol{T}_{1i}$ is given by (5.6), and

$$\hat{\boldsymbol{B}}_{2j}(d) = \boldsymbol{T}_{2j}^{-1} \sum\nolimits_{s_{2j}} \frac{\tilde{w}_{1k} w_{2k} z_k y_k(d)}{C_{2k}} \qquad (6.6)$$

with $\boldsymbol{T}_{2j}$ given by (5.10), and

$$\hat{y}_{2k}(d) = \boldsymbol{z}_k' \hat{\boldsymbol{B}}_{2j}(d) \text{ for } k \in (s_{ij} \cap U_d).$$

Remark 6.1: Note that for each new domain of interest, the variance estimator (6.2) requires two new sets of domain dependent residuals, $e_{1k}(d)$ and $e_{2k}(d)$. Moreover, these are required for *all* of the units k in the second-phase sample s_2, including units outside the domain. Variance estimation for domains can therefore be cumbersome.

Remark 6.2: In practice the computation of estimated variances is seldom carried out as a double sum. For some important designs, the double sums reduce, after some algebraic manipulation, to single sum expressions. Examples of this occur for single sampling and for stratified single random sampling in both phases. Explicit algebraic developments for the variances have been given the former case by Särndal *et al.* (1992), and in the later case by Hidiroglou (1995), and Binder, Babyak, Brodeur, Hidiroglou and Jocelyn (1997).

7. APPLICATIONS WITH POSTSTRATIFICATION AT THE FIRST PHASE

7.1 The Case of the Tax Sample at Statistics Canada

An application of the calibration group approach in section 5 has been in use at Statistics Canada, in the two-phase design for sampling of tax records. The example is important because it provides the extension to two-phase designs of the traditional postratification technique as used in a single phase design. The sampling procedure, the post-stratification criteria, and the estimators are described in Armstrong and St-Jean (1994). We now show how these estimators are obtained as special case of the technique in section 5. The sampling design, in each phase, is stratified Bernouilli, carried out with the permanent random number technique. The two stratifications are based on different criteria. The realized sample sizes are random at each phase on account of the Bernouilli sampling. To offset the resulting tendency toward an increased variance, poststratification is carried out at both phases of sampling. The two poststratification criteria are different. We have in effect two crossing poststratifications. In the terminology of section 5, the first phase poststrata are the first-phase calibration groups. They are denoted as U_i ; $i = 1, ..., I$, and the group membership of a unit k is indicated by the vector by Δ_{1k} given by (5.1). The second phase poststrata are the second phase calibration groups. They are denoted as $s_{1j}, j = 1, ..., J$ and the corresponding membership of a unit k is indicated by the vector Δ_{2k} given by (5.3).

The first-phase calibration is carried out using the information about the first-phase poststrata sizes, N_i. In this survey design, there is no supplementary information, so $z_{1k} = 1$ for all k in (5.5), yielding $\boldsymbol{x}_{1k} = \Delta_{1k}$. Specifying $C_{1k} = 1$ for all k we obtain from (5.7) that

$$g_{1k} = N_i / \hat{N}_{1i} \qquad (7.1)$$

for all $k \in s_{1i}$ where $\hat{N}_{1i} = \sum_{s_{1i}} w_{1k}$ estimates the known first-phase poststratum count N_i, and $s_{1i} = s_i \cap U_i$ denotes the part of the first-phase sample s_1 that falls in the first-phase poststratum U_i.

We arrive at the estimator of Armstrong and St-Jean (1994) by carrying out the second-phase calibration with $\boldsymbol{x}_k = \Delta_{2k}$, that is, we have $z_k = 1$ for all k in (5.8). This is a reduced $\boldsymbol{x}_k$-vector specification since it does not involve $\boldsymbol{x}_{1k}$. Specifying $C_{2k} = 1$ for all $k \in s_{1i}$, and using (5.9) and (3.10), we obtain the overall calibrated weights

$$g_k^* = \frac{N_i}{\hat{N}_{1i}} \frac{\hat{N}_{1j}}{\hat{N}_{2j}} \qquad (7.2)$$

for all $k \varepsilon s_{2ij}$, where

$$\hat{N}_{1j} = \sum_{i=1}^{I} \left(\frac{N_i}{\hat{N}_{1i}} \right) \hat{N}_{1ij} ; \hat{N}_{2j} = \sum_{i=1}^{I} \left(\frac{N_i}{\hat{N}_{1i}} \right) \hat{N}_{2ij} \qquad (7.3)$$

with $\hat{N}_{1ij} = \sum_{s_{1ij}} w_{1k}$ and $\hat{N}_{2ij} = \sum_{s_{2ij}} w_k^*$. Here, $s_{2j} = s_2 \cap s_{1j}$ denotes the part of the second-phase sample s_2 that falls in the second-phase poststratum s_{1j}, and $s_{1ij} = U_i \cap s_{1j}$; $s_{2ij} = s_2 \cap U_i \cap s_{1j}$. It follows that the estimator of the total $Y(d)$ for a given domain U_d is given by $\hat{Y}(d) = \sum_{s_2} w_k^* g_k^* y_k(d)$, or equivalently as

$$\hat{Y}(d) = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{N_i}{\hat{N}_{1i}} \frac{\hat{N}_{ij}}{\hat{N}_{2j}} \sum_{s_{2ij}} w_k^* y_k(d).$$

The estimated variance requires two types of residuals that are easily obtained from the general expressions given in Section 6.

Alternatives exist to the reduced vector specification $\boldsymbol{x}_k = \Delta_{2k}$ used for this design. We therefore examine what the estimator would look like under a full vector specification. For the first-phase calibration, as earlier, let $\boldsymbol{x}_{1k} = \Delta_{1k}$ corresponding to $z_{1k} = 1$ for all k in (5.8). The first-phase g-factors g_{1k} are then given by (7.1). In this survey, information is available for assigning every unit $k \in s_1$ to one of the $I \times J$ cells formed by cross-classifying the two poststratification criteria. Therefore, the vector $\boldsymbol{x}_k$ for the second-phase calibration can be taken as

$$\boldsymbol{x}_k' = \Delta_{1k} \otimes \Delta_{2k}' \qquad (7.4)$$

This is a full vector specification in that it includes the first-phase information carrier Δ_{1k}. Let us also specify $C_{2k} = 1$ for all k. Since (7.4) is of the form (5.8), the second-phase g-factors g_{2k} are obtainable group-by-group from (5.9) with $z_k = \Delta_{1k}$. The overall calibration factors are given by

$$g_k^* = \frac{N_i}{\hat{N}_{1i}} \frac{\hat{N}_{1ij}}{\hat{N}_{2ij}} \qquad (7.5)$$

for all $k \in s_{2ij}$. Here, $\hat{N}_{1i}$ is defined in (7.1), and $\hat{N}_{1ij}$ and $\hat{N}_{2ij}$ are as in (7.3). These overall calibration factors are the product of two poststratified calibration factors. They are all positive and well defined, provided all sample cells s_{2ij} are non-empty. Collapsing of small cells s_{2ij} with relatively large non-empty cells is recommended for stable estimation. As pointed out in Remark 3.4, the overall weights obtained from (7.5) reproduce the known first-phase postrata sizes N_i, whereas those obtained from (7.2) do not.

Remark 7.1: Let us compare the calibration factors (7.2) and (7.5), resulting, respectively, from the reduced form $\boldsymbol{x}_k = \Delta_{2k}$ and from the full form (7.4). Both factors are a product of two terms. The only difference lies in the second term. In both cases, the computation of the second term requires cross-classification information. That is, for every $k \in s_1$, we need to identify the cross-classification cell ij to which k belongs. In the case of the reduced vector, the cell information is pooled across the first-phase groups. For the full vector, the cell information is kept separate, and one would expect the resulting weights to be more efficient.

Remark 7.2: For the second-phase calibration, an alternative to (7.4) that also captures the information about the first-phase poststrata is to use

$$\boldsymbol{x}_k' = \left(\Delta_{1k}', \Delta_{2k}' \right). \qquad (7.6)$$

Note that with this specification, there is only one calibration group in the second phase, namely the whole first-phase sample s_1.

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

The Survey on Employment Payrolls, and Hours (SEPH) covers all sectors of Canadian industry, and collects data on four principal variables: (i) salaries and payments to employees (denoted as z_2; called payrolls); (ii) number of employees (z_3; employment); (iii) hours worked by employees (y_1; hours); and (iv) summarized earnings (y_2; earnings).

SEPH (1994) 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 within strata ranging from 10% to 100%. The strata are defined by region. A region is made up of one or more Canadian provinces. We describe the estimation for SEPH by considering one specific region.

For units selected in the first-phase sample, two variables are transcribed, namely, payrolls (z_2) and number of employees (z_3). In the second-phase, a simple random sample is drawn. Data on the two variables of interest, y_1 and y_2, are collected for respondents in this sample. In addition, classification by industry and province is recorded for sampled units. The first-phase sample is poststratified by employment size groups. These are used as first-phase calibration groups and denoted U_i; $i = 1, ..., I$. Their sizes denoted as N_i for $i = 1, ..., I$ are assumed known. The vector $\boldsymbol{x}_{1k}$ used for a first-phase calibration is of the form (5.5), where Δ_{1k} is given by (5.1) and $z_{1k} = 1$ for all k. We choose $C_{1k} = 1$ for all k. It follows from (5.7) that the first-phase g-factors are

$$g_{1k} = N_i / \hat{N}_{1i} \tag{7.7}$$

for all $k \in s_{1i} = s_i \cap U_i$, where $\hat{N}_{1i} = \sum_{s_{1i}} w_{1k}$, $i = 1, ..., I$.

We now turn to second-phase calibration. It is carried out using calibration groups $s_{1j}, j = 1, ..., J$, identified by the vector Δ_{2k} given by (5.3). These groups are based on a province by industry classification. They are constructed so that: (i) there is a strong regression relationship between y_k and the two z-variables, and that (ii) there are at least 30 observations within each group. The $J(I+2)$ dimensional $\boldsymbol{x}_k$-vector for the second-phase calibration is given by

$$\boldsymbol{x}_k' = \Delta_{2k}' \otimes (\Delta_{1k}', z_{2k}, z_{3k}) \tag{7.8}$$

This specification requires (see Table 1) that every $k \in s_1$ can be classified into one of the I by J cells formed by crossing the calibration groups in the two phases. Let $s_{2j} = s_2 \cap s_{1j}$; $s_{1ij} = s_{1j} \cap U_i$; $s_{2ij} = s_2 \cap s_{1ij}$. Also, the quantitative variable values z_{2k} (payrolls) and z_{3k} (number of employees) must be known for $k \in s_1$. The $\boldsymbol{x}_k$-vector specification given by (7.8) is full, because it incorporates $\boldsymbol{x}_{1k} = \Delta_{1k}$. A reduced vector, ignoring the first-phase groups, would be $\boldsymbol{x}_k' = \Delta_{2k}' \otimes (z_{2k}, z_{3k})$.

As in Example 7.1, we have two crossing sets of calibration groups.

Since the $\boldsymbol{x}_k$-vector (7.8) has the structure defined by (5.8), we used (5.9) to derive the second-phase g-factors for each group $j = 1, ..., J$. It follows from (7.8) that we are fitting, within each second-phase calibration group, a separate regression of y_k on $\zeta_k = (z_{2k}, z_{3k})'$ with an intercept that varies with the first-phase calibration group.

Specifying $C_{2k} = 1$ for all k, and using the additive form, $g_k^* = g_{1k} + g_{2k} - 1$, for the overall calibration factors, we obtain after some algebra

$$g_k^* = G_{1i} G_{2ij} + \boldsymbol{H}_j' \boldsymbol{T}_j^{-1} \left(\zeta_k - \bar{\zeta}_{s_{2ij}} \right)$$

for all $k \in s_{2ij}$, where

$$G_{1i} = N_i / \hat{N}_{1i},\ G_{2ij} = \hat{N}_{1ij} / \hat{N}_{2ij}$$

$$\boldsymbol{H}_j = \sum_{i=1}^{I} \hat{N}_{1ij} G_{1i} \left(\bar{\zeta}_{s_{1ij}} - \bar{\zeta}_{s_{2ij}} \right)$$

$$\boldsymbol{T}_j = \sum_{i=1}^{I} \sum\nolimits_{s_{2ij}} w_k^* \left(\zeta_k - \bar{\zeta}_{s_{2ij}} \right) \left(\zeta_k - \bar{\zeta}_{s_{2ij}} \right)'$$

with

$$\bar{\zeta}_{s_{1ij}} = \sum\nolimits_{s_{1ij}} \frac{w_{1k} \zeta_k}{\hat{N}_{1ij}};\ \bar{\zeta}_{s_{2ij}} = \sum\nolimits_{s_{2ij}} \frac{w_k^* \zeta_k}{\hat{N}_{2ij}};\ \hat{N}_{1ij} = \sum\nolimits_{s_{1ij}} w_{1k};$$

and $\hat{N}_{2ij} = \sum_{s_{2ij}} w_k^*$.

It follows that we can write the estimator (6.1) as $\hat{Y}(d) = \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{Y}_{ij}(d)$ with

$$\hat{Y}_{ij}(d) = G_{1i} \hat{N}_{1ij} \{ \bar{y}_{s_{2ij}}(d) + (\bar{\zeta}_{s_{1ij}} - \bar{\zeta}_{s_{2ij}})' \hat{\boldsymbol{B}}_j(d) \}$$

where

$$\bar{y}_{s_{2ij}}(d) = \sum\nolimits_{s_{2ij}} w_k^* y_k(d) / \hat{N}_{2ij}$$

and $\hat{\boldsymbol{B}}_j(d) = \boldsymbol{T}_j^{-1} \sum_{i=1}^{I} \sum_{s_{2ij}} w_k^* (\zeta_k - \bar{\zeta}_{s_{2ij}}) y_k(d)$.

The form of $\hat{Y}(d)$ is easy to understand. It is composed of $I \times J$ cell estimates $\hat{Y}_{ij}(d)$, each reflecting the regression of $y_k(d)$ on ζ_k. Note that the two-dimensional slope vector $\hat{\boldsymbol{B}}_j(d)$ is obtained by pooling data across the first-phase groups. This is because the specification (7.8) of $\boldsymbol{x}_k$ allows the intercept, but not the two regression slopes, to vary with the first-phase groups.

8. CONCLUSIONS

Two-phase designs have the advantage of being both economical and efficient. The present paper has provided a general theory for such designs when auxiliary information is present in each phase.

Our goal is to incorporate this two-phase survey methodology into Statistics Canada's Generalized Estimation System (GES) described in Estevao *et al.* (1995). The GES is a general purpose program that currently handles domain estimation for arbitrary single phase designs and incorporates auxiliary information in its estimation process. In this paper we have extended the basic principles of the GES, including the important idea of calibration groups, to two-phase designs.

We have illustrated the theory by showing its use in two current surveys at Statistics Canada. Given its generality, the theory has potential application to any two-phase sample design that uses auxiliary information.

REFERENCES

ARMSTRONG, J., and ST-JEAN, H. (1994). Generalized regression estimation for a two-phase sample of tax records. *Survey Methodology*, 20, 97-106.

BINDER, D.A. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22, 17-22.

BINDER, D.A., BABYAK, C., BRODEUR, M., HIDIROGLOU, M.A., and JOCELYN, W. (1997). Variance Estimation for Two-phase Stratified Sampling. Contributed paper presented at the Annual Meeting of the American Statistical Association, Los Angeles.

BREIDT, J., and FULLER, W.A. (1993). Regression weighting for multiphase samples. *Sankhyā*, 55, 297-309.

CHAUDHURI, A., and ROY, D. (1994). Model assisted survey sampling strategy in two phases. *Metrika*, 41, 355-362.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd ed.). New York: John 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, 1013-1020.

DUPONT, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, 21, 125-136.

ESTEVAO, V., 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.

HIDIROGLOU, M.A. (1995). Sampling and estimation for stage one of the Canadian Survey of Employment, Payrolls and Hours redesign. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 123-128.

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 of the 1995 Annual Research Conference*. U.S. Bureau of the Census, 171-197.

HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 873-878.

NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *Journal of The American Statistical Association*, 33, 101-116.

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.C., and MOHL, C.A. (1996). Understanding calibration estimation in survey sampling. *Survey Methodology*, 22, 107-115.

STUKEL, D., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1996). Variance estimation for calibration estimators: a comparison of jackknifing versus linearization. *Survey Methodology* , 22, 117-125.