TWO-PHASE SAMPLING OF TAX RECORDS FOR BUSINESS SURVEYS

John Armstrong, Clayton Block and K.P. Srinath, Statistics Canada
John Armstrong, 11-N RH Coats Bldg, Ottawa, Ontario K1A 0T6

KEY WORDS: optimal allocation, convex programming, efficiency

1. Introduction

In the last few years, there has been an increase in the use of administrative records by statistical agencies. This is in response to demands for more detailed data and requirements to reduce costs as well as the burden imposed on respondents, especially small businesses. Income tax files are one source of annual data. A new strategy for the collection and integration of economic data is being implemented at Statistics Canada. According to the strategy, annual economic data for large businesses are collected through mail-out sample surveys and data for small businesses are obtained from a sample of tax records. Estimates of financial variables for the business population are obtained by combining estimates from the two sources.

This paper describes the sample design adopted for the sampling of tax records. There is a requirement to produce estimates for cells defined by four-digit Standard Industrial Classification (SIC) code and province. The tax records in the population can be stratified much more accurately by two-digit SIC (SIC2) and province than by four-digit SIC (SIC4). Therefore, a two-phase sampling procedure is used in order to sample by SIC4. It is necessary to determine first- and second-phase sampling fractions that minimize cost subject to constraints on the coefficients of variation of each SIC4 estimate of gross business income at the province level. This optimal allocation problem differs from the allocation problems for two-phase designs described by Cochran (1977) and Smith (1989).

A description of the two-phase sample design is given in section two. Methods used for estimation are described in section three. The important problem of sample allocation is considered in section four. The efficiency of the two-phase design relative to one-phase alternatives is examined in section five. Some conclusions are given in the sixth section.

2. Sampling Design

2.1 Background

The target population for tax sampling is the population of businesses with gross income over $25,000, excluding large businesses that are covered by mail-out sample surveys. Revenue Canada provides Statistics Canada with information concerning taxfilers. There are two types of taxfilers, T1s and T2s. A T1 taxfiler is an individual, who may own all or part of one or more unincorporated businesses, while a T2 taxfiler is an incorporated business. For tax year 1988, there were 466,000 T2 taxfilers in the target population and 533,000 T1 taxfilers with ownership shares in one or more in-scope businesses. About 95.5% of these T1 taxfilers had an ownership share in only one business. About 88% of in-scope businesses owned by T1 taxfilers were owned by a single taxfiler and the rest were partnerships.

During the early and mid-1980's, Statistics Canada selected an annual sample of tax returns using a one-phase sample design with stratification by geographic region and size (gross income). The precision of estimates at the SIC4 *x* province level varied from one industry to another. It was considered desirable to introduce sampling by SIC4 in order to standardize the precision of the estimates. Revenue Canada attempts to assign an SIC code with four digits to all T2 taxfilers as well as to most T1 taxfilers reporting business income. It was decided not to use SIC4 codes assigned by Revenue Canada during sample selection for two reasons. First, studies indicated that the accuracy of SIC coding based exclusively on information provided to Revenue Canada was relatively low. For example Estevao *et al* (1986) studied the accuracy of SIC codes based on Revenue Canada information using a sample of 1984 tax returns filed by incorporated businesses with revenue of less than $10 million. Only about 54% of businesses in their test sample were assigned a correct SIC4. Second, coding operations at Revenue Canada are not under the control of Statistics Canada and there was no guarantee that Revenue Canada would continue to code all four digits of SIC. At the same time, the first two digits of SIC codes assigned by Revenue Canada were considered sufficiently accurate and there were no doubts that two-digit coding of tax returns would continue. Consequently, a two-phase sample design was introduced.

2.2 First-Phase Sample Selection

The first-phase sample is a sample of taxfilers selected at Revenue Canada using strata defined using SIC2, province and size. (For each T1 taxfiler, the total income of all businesses wholly or partially owned by the taxfiler is used as the size measure.) The first-phase sample is a longitudinal sample. All taxfilers that are included in the first-phase sample in tax year *Y* (TY(*Y*)) and are still in the population for purposes of tax sampling in TY(*Y+1*) are included in the first-phase sample for TY(*Y+1*). Taxfilers may be added to the first-phase sample each year in order to improve the precision of certain estimates and to replace taxfilers sampled in previous years that are no longer in the target population.

Selection of the first-phase sample is done using Bernoulli sampling. Each taxfiler is assigned a pseudo-random number (hash number) in the interval (0,1) generated by a hashing function that uses the unique taxfiler identifier as input. The random number assigned a given taxfiler does not change from one tax year to the next. Denote the SIC2 codes used to define first-phase sampling strata within a province by $e = 1, 2, \ldots, E$ and denote size groups by $q = 1, 2, \ldots, Q$. Let υ_{eq} denote the first-phase sampling fraction for first-phase stratum *eq*, corresponding to SIC2 code *e* and size group *q*, for TY(*Y*).

Let R_i denote the pseudo-random number associated with taxfiler *i* and suppose that taxfiler *i* falls in first-phase stratum *eq* in TY(*Y*). If taxfiler *i* was in the target population for TY(*Y-1*) the taxfiler may have fallen in a different stratum in TY(*Y-1*) since the industrial activity and size of a business may change between tax years. If taxfiler *i* is not in the first-phase sample

of TY(Y-1), taxfiler *i* is selected for the first-phase sample in TY(Y) if $0 \le R_i < \upsilon_{sq}$. Since taxfiler identifiers do not change over time, Bernoulli sampling facilitates selection of a longitudinal sample. Sample sizes obtained using this method are random variables. First-phase selection probabilities are updated each year to reflect the longitudinal nature of the first-phase sample. Details of the updating method are available in Armstrong, Block and Srinath (1991).

2.3 Second-Phase Sample Selection

Let $J = \{j\}$ denote the population of businesses that is the target population for tax sampling. In order to select the second-phase sample, statistical entities are created using information about businesses corresponding to taxfilers in the first-phase sample. T1 tax returns include income and expense data for all businesses wholly or partially owned by the taxfiler, as well as ownership percentage information. A statistical entity, denoted by *(i,j)*, is created for every taxfiler-business combination in the first-phase sample. A new SIC code, considered accurate to four digits, is assigned to all statistical entities by Statistics Canada.

Conceptually, the second-phase sample is a sample of businesses. Operationally, it is a sample of taxfilers selected using statistical entities. If one statistical entity corresponding to a T1 taxfiler is selected for the second-phase sample then all statistical entities corresponding to the taxfiler are selected. Strata defined at the level of SIC4 *x* province *x* size are used during selection of the second-phase sample. The total revenue of business *j* is used as the size variable for statistical entity *(i,j)*. Let J_i be a set of indices identifying businesses that are wholly or partially owned by taxfiler *i*. Denote the SIC4 codes within a province by $f = 1, 2, \ldots, F$, and let υ'_{fq} denote the second-phase sampling fraction in stratum *fq* for TY(Y). Suppose that statistical entity *(i,j)* falls in second-phase stratum *fq(j)* in TY(Y). The probability that statistical entity *(i,j)* will be selected in the second-phase sample for TY(Y) is

$$p2_{ij} = p2_i = 1 - \prod_{j' \in J_i} \left(1 - \upsilon'_{fq(j')}\right).$$

Business *j* will be included in the second-phase sample if a statistical entity involving the business is selected.

The second-phase sample is selected using Bernoulli sampling. Although the second-phase sample is not a longitudinal sample, the number of statistical entities in common between second-phase samples for consecutive tax years can be controlled by varying the overlap of hash intervals. Data for approximately thirty-five financial variables is captured for each business in the second-phase sample.

3. Estimation

Methods used for estimation are described in this section. A post-stratified version of the Horvitz-Thompson estimator is employed. The approach used to adjust for the effects of partnerships is described in subsection 3.1. Subsection 3.2 contains details of the post-stratified ratio adjustment.

3.1 Partnership Adjustment

The second-phase selection probability for each taxfiler is calculated using information about the associated statistical entities. In order to construct estimates for the population of businesses based on this sample, an adjustment for the effects of partnerships is required. If business *j* is a partnership, it will be included in the second-phase sample if any of the corresponding taxfilers are selected. The usual Horvitz-Thompson estimator must be adjusted for partnerships to avoid over-estimation. Let δ_{ij} denote the proportion of business *j* owned by taxfiler *i* and suppose that taxfiler *i* is selected for the second-phase sample. The data for business *j* is adjusted by multiplying it by δ_{ij} so that only the component of income and expense items corresponding to taxfiler *i* is included in estimates. Note that adjusted data for business *j* is used only during tabulation of estimates and is not used during sample allocation or selection.

Let y_j denote the value of the variable *y* for business *j*. The Horvitz-Thompson estimator of the total of *y* over domain *d*, incorporating adjustment for partnerships, is given by

$$\hat{Y}(d) = \sum_{i \in s2} \sum_{j \in J_i} \delta_{ij} \cdot y_j(d) / (P1_i \cdot P2_i),$$

where $y_j(d) = y_j$ if business *j* falls in domain *d* and is otherwise zero.

3.2 Post-Stratified Ratio Adjustment

Sunter (1986) shows that, in the case of a one-phase design using Bernoulli sampling, the estimator analogous to $\hat{Y}(d)$ has a large variance. He also considers a ratio form of the estimator, adjusted for differences between actual and expected sample sizes. He notes that the ratio form has a small bias and a variance that is considerably smaller than the unadjusted version. The methodology used to produce tax estimates was proposed by Choudhry, Lavallée and Hidiroglou (1989). Ratio adjustments to account for discrepancies between actual and expected sample sizes are applied within post-strata during weighting of both the first- and second-phase samples.

Let $U = \{u\}$ denote a set of first-phase post-strata and suppose that post-stratum *u* contains N_u taxfilers. The estimate of the number of taxfilers in the population that fall in first-phase post-stratum *u*, based on the first-phase sample, is

$$\hat{N}_u = \sum_{i \in s1 \cap u} (1 / p1_i).$$

The post-stratified first-phase weight for taxfiler $i, i \in u$, is

$$W1_i = (1 / p1_i) \cdot (N_u / \hat{N}_u).$$

Similarly, let $V = \{v\}$ define a set of second-phase post-strata. The estimate of the number of taxfilers in second-phase post-stratum *v*, based on the first-phase sample, is

$$\hat{N}_v = \sum_{i \in s1 \cap v} W1_i.$$

An alternative estimate, using only units in the second-phase sample, is

$$\tilde{N}_v = \sum_{i \in s2 \cap v} W1_i / p2_i.$$

The post-stratified second-phase weight for statistical entities corresponding to taxfiler *i*, is

$$W2_i = (1 / p2_i) \cdot (\hat{N}_v / \tilde{N}_v).$$

The post-stratified estimate of the total of *y* over domain *d* is given by

$$\hat{Y}_{PS}(d) = \sum_{i \in s2} \sum_{j \in J_i} \delta_{ij} \cdot W1_i \cdot W2_i \cdot y_j(d). \qquad (1)$$

Using standard methods for approximating the variance of a ratio estimate, one can show that the variance of $\hat{Y}_{PS}(d)$ is approximately given by

$$V(\hat{Y}_{PS}(d)) \approx \sum_{u} \sum_{i \in u} \frac{(1-p1_i)}{p1_i} \cdot \left(y_i(d) - \frac{Y_u(d)}{N_u(d)} \right)^2 + \sum_{v} \sum_{i \in v} \frac{(1-p2_i)}{p1_i \cdot p2_i} \cdot \left(y_i(d) - \frac{Y_v(d)}{N_v(d)} \right)^2, \quad (2)$$

where

$$y_i(d) = \sum_{j \in J_i} \delta_{ij} \cdot y_j(d),$$

$Y_u(d)$ and $Y_v(d)$ are population totals for the variable y over the portions of the domain d belonging to post-strata u and v respectively, $N_u(d)$ is the number of taxfilers in the portion of domain d belonging to post-stratum u, and $N_v(d)$ is the number of taxfilers belonging to domain d as well as post-stratum v. Taxfiler i belongs to domain d if one of the corresponding businesses belongs to the domain. A closed form estimator of (2) in given in Choudhry, Lavallée and Hidiroglou (1989).

4. Sample Allocation

In this section, methods used to allocate the tax sample are described. Precision requirements for estimates of gross business income are specified for each SIC4 x province domain. It is necessary to determine an allocation that is optimal in the sense that sampling costs are minimized while all precision constraints are satisfied.

The optimal allocation for a two-phase sample design when there is only one precision constraint, involving an overall estimate, is well-known and is described by Cochran (1977). Recently Smith (1989) studied a problem involving allocation of a sample of fixed size to minimize a loss function involving a number of domain estimates. The current problem differs because the optimal allocation must satisfy one precision constraint for each SIC4 x province domain.

The assumptions used to permit a mathematically concise formulation of the optimal allocation problem are mentioned in subsection 4.1. The problem is formulated in subsection 4.2 and a method of obtaining the exact solution is described. A method that provides an approximate solution is discussed in subsection 4.3. Exact and approximate methods are compared in Armstrong and Le Petit (1991) using data from the province of Québec for TY(1988). The results of the comparison indicate that efficiency losses due to use of the approximate method are relatively small.

4.1 Background

To obtain variance estimates for use in sample allocation, it is assumed that SIC4 codes used during selection of the second-phase sample are always consistent with SIC2 codes assigned by Revenue Canada. This is not always true for three reasons. First, in the case of a T1 taxfiler with more than one business, the SIC2 code assigned to the taxfiler by Revenue Canada corresponds to the business with the largest revenue. Second, most SIC2 codes assigned by Revenue Canada reflect the current economic activity of taxfilers. The SIC4 code assigned to a business and used to select the second-phase sample is based on activity during a tax year in which either the corresponding taxfiler was selected for the first time for the first-phase sample or the code was updated using a process independent of second-phase sample selection. Finally, coding errors can lead to inconsistencies.

Consider one SIC2 x province cell containing N units. Each unit falls into one of H SIC4 x province cells. The SIC2 x province cell contains G first-phase SIC2 x province x size strata, and $G \cdot H$ second-phase SIC4 x province x size strata. Let N_g denote the number of units in the SIC2 x province x size stratum g, and let υ_g be the corresponding first-phase sampling fraction. Similarly, let N_{gh} denote the number of units, and υ_{gh} the second-phase sampling fraction, respectively, for the SIC4 x province x size stratum gh. Finally, let $Y(h)$ denote the population total of gross business income for SIC4 x province cell h. Conditional on sample size, Bernoulli sampling is equivalent to simple random sampling when all units in each stratum have the same probability of selection. The variance of $\hat{Y}_{2\text{-}PH}(h)$, the Horvitz-Thompson estimate of $Y(h)$ for a two-phase design using simple random sampling, is

$$V_h = \sum_{g} \left(\frac{1}{\upsilon_g \cdot \upsilon_{gh}} - 1 \right) \cdot A_{gh} + \sum_{g} \left(\frac{1}{\upsilon_g} - 1 \right) \cdot B_{gh}, \quad (3)$$

where

$$A_{gh} = N_{gh} \cdot S^2_{gh},$$

$$B_{gh} = \left(\frac{N_g - N_{gh}}{N_g - 1} \right) \cdot \left(\frac{Y^2_{gh}}{N_{gh}} - S^2_{gh} \right),$$

and Y^2_{gh} and S^2_{gh} are the total and variance of y, respectively, for second-phase stratum gh. Details of the derivation of (3) are available in Armstrong, Block and Srinath (1991).

The variance given by (3) is used during sample allocation. It is assumed that there is a one-to-one correspondence between taxfilers and businesses. Let $\hat{Y}_{PS}(h)$ denote the post-stratified Horvitz-Thompson estimator of the population total of y for SIC4 x province domain h. The variance of $\hat{Y}_{PS}(h)$ is given by (2), replacing d by h. Note that this variance is close to (3) if sampling strata are used as post-strata and all units in each sampling stratum have the same probability of selection.

4.2 Optimal Allocation

Let K_1 and K_2 denote the costs of selecting a unit at the first and second phases of sampling, respectively. These costs do not vary between strata. The optimal allocation problem is the problem of minimizing the cost function

$$F = K_1 \cdot \sum_{g} \upsilon_g \cdot N_g + K_2 \cdot \sum_{g} \sum_{h} \upsilon_g \cdot \upsilon_{gh} \cdot N_{gh}, \quad (4)$$

with respect to υ_g, $g = 1, 2, \ldots, G$, and υ_{gh}, $g = 1, 2, \ldots, G$, $h = 1, 2, \ldots, H$, under the constraints

$$\sum_{g} \left(\frac{1}{\upsilon_g \cdot \upsilon_{gh}} - 1 \right) \cdot A_{gh} + \sum_{g} \left(\frac{1}{\upsilon_g} - 1 \right) \cdot B_{gh} \le C_h^2 \cdot Y_h^2, \quad \forall h,$$

$$0 < \upsilon_g \le 1, \quad \forall g, \quad (5)$$

$$0 < \upsilon_{gh} \le 1, \quad \forall g, h,$$

where C_h denotes the target coefficient of variation for SIC4 x province cell h.

Wu (1989) suggests a simplification of the problem by dividing it into two parts that can be solved iteratively. Let $\upsilon_g^{(r)}$ and $\upsilon_{gh}^{(r)}$ denote the estimates of the optimal values of υ_g and υ_{gh} obtained after r iterations.

Each iteration includes the following steps:

(i) Minimize

$$F = \sum_g \left(N_g + \frac{K_2}{K_1} \sum_h \upsilon_{gh}^{(r-1)} \cdot N_{gh} \right) / (X_g^{(r)} + 1)$$

with respect to $X_g^{(r)}$, $g = 1, 2, \ldots, G$, subject to the constraints

$$C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{X_g^{(r)} + 1}{\upsilon_{gh}^{(r-1)}} - 1 \right) \cdot A_{gh} - \sum_g X_g^{(r)} \cdot B_{gh} \geq 0, \quad \forall h$$

$$X_g^{(r)} \geq 0, \quad \forall g.$$

(ii) Calculate $\upsilon_g^{(r)} = 1/(X_g^{(r)} + 1)$, $\forall g$,

and minimize, independently for each h,

$$F_h = \sum_g \upsilon_g^{(r)} \cdot \upsilon_{gh}^{(r)} \cdot N_{gh}$$

with respect to $\upsilon_{gh}^{(r)}$, $g = 1, 2, \ldots, G$ subject to the constraints

$$C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{\upsilon_g^{(r)} \cdot \upsilon_{gh}^{(r)}} - 1 \right) \cdot A_{gh} - \sum_g \left(\frac{1}{\upsilon_g^{(r)}} - 1 \right) \cdot B_{gh} \geq 0,$$

$$0 < \upsilon_{gh}^{(r)} \leq 1, \quad \forall g.$$

Reasonable starting values for the first iteration are $\upsilon_{gh}^{(0)} = 1$, $g = 1, 2, \ldots, G$, $h = 1, 2, \ldots, H$. At step (i), the transformation of variables given by $X_g = 1/\upsilon_g - 1$ is employed. Since the cost function to be minimized is a concave function of X_g, $g = 1, 2, \ldots, G$, and the constraints are convex functions, the optimization problem is a convex programming problem. To find the global solution of a convex programming problem it is sufficient to find a local solution. Note that only G variables are involved, while there are $G \cdot (H + 1)$ variables in the optimization problem given by (4) and (5). It will be shown in subsection 4.3 that the solution of step (ii) has a closed form.

4.3 Approximate Method

An approximation to optimal allocation is used in practice. Assuming that all second-phase sampling fractions are equal to one, an approximation to the optimal allocation of the first-phase sample is calculated. Then the second-phase sample is allocated, conditional on the first-phase sampling fractions. Since the cost of sampling a unit in both phases of sampling does not depend on the stratum in which the unit falls, minimizing cost is equivalent to minimizing sample size at each step of this method.

An approximate solution to the optimal allocation problem for a one-phase sample design can be obtained by finding the minimum, independently for each h, of

$$F^{(h)} = \sum_g \upsilon_{g|h} \cdot N_g \tag{6}$$

with respect to $\upsilon_{g|h}$, $g = 1, 2, \ldots, G$, subject to the constraints

$$\sum_g \left(\frac{1}{\upsilon_{g|h}} - 1 \right) \cdot (A_{gh} + B_{gh}) \leq C_h^2 \cdot Y_h^2, \tag{7}$$

$$0 < \upsilon_{g|h} \leq 1, \quad \forall g.$$

The minimum of (6) is obtained when (7) holds with equality. The solution to the minimization problem defined by (6) and (7) can be obtained in a straight-forward manner using the method of Lagrange and is given by

$$\upsilon_{g|h} = ((A_{gh} + B_{gh})/N_g)^{1/2} \cdot \sum_g ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} /$$

$$\left(C_h^2 \cdot Y_h^2 + \sum_g (A_{gh} + B_{gh}) \right). \tag{8}$$

If one or more of the sampling fractions given by (8) are greater than one, one can set them equal to one and solve a modified allocation problem with a reduced number of strata. This approach corresponds to the overallocation procedure discussed by Cochran (1977). Finally, υ_g is set equal to the largest value in the set $\{\upsilon_{g|h}, \; h = 1, 2, \ldots, H\}$ for $g = 1, 2, \ldots G$.

Given first-phase sampling fractions, optimal second-phase sampling fractions can be easily determined. Let n_{gh} denote the number of units slected in the first-phase sample that fall into second-phase stratum gh. Note that n_{gh} will not necessarily be equal to its expected value, given by $\upsilon_g \cdot N_{gh}$. Assume that, for the SIC4 x province cell h, the size strata included in the allocation problem correspond to a set of integers, Γ. We set the second-phase sampling fractions equal to one for those size strata that are not included in the allocation problem. The problem of allocating the second-phase sample is equivalent to the problem of finding the minimum of

$$F_h = \sum_{g \in \Gamma} \upsilon_{gh} \cdot n_{gh} \tag{9}$$

with respect to υ_{gh}, $g \in \Gamma$, subject to the constraints

$$\sum_{g \in \Gamma} \frac{1}{\upsilon_g} \cdot \left(\frac{1}{\upsilon_{gh}} - 1 \right) \cdot A_{gh} \leq M_h, \tag{10}$$

$$0 < \upsilon_{gh} \leq 1, \quad g \in \Gamma, \tag{11}$$

where

$$M_h = C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{\upsilon_g} - 1 \right) \cdot (A_{gh} + B_{gh}).$$

It is easy to show that (9) attains a minimum when (10) holds with equality. Using the method of Lagrange and ignoring (11), one obtains

$$\upsilon_{gh} = [A_{gh}/(\upsilon_g \cdot n_{gh})]^{1/2} \cdot$$

$$\sum_{g \in \Gamma} (n_{gh} \cdot A_{gh}/\upsilon_g)^{1/2} / D_{\Gamma h}, \tag{12}$$

where

$$D_{\Gamma h} = C_h^2 \cdot Y_h^2 + \sum_{g \in \Gamma} \left(\frac{1}{\upsilon_g} \right) \cdot A_{gh} - \sum_g \left(\frac{1}{\upsilon_g} - 1 \right) \cdot (A_{gh} + B_{gh}).$$

Note that there is no solution to the allocation problem unless D_{rh} is positive. If certain values of υ_{gh} are greater than one, overallocation can be used. In practice, N_{gh}, Y_h and S^2_{gh} are, of course, unknown, and estimates must be employed. Note that if we replace n_{gh} in (12) by its expectation with respect to the first phase of sampling, $\upsilon_g \cdot N_{gh}$, the closed form solution of the minimization problem involved in the second step of the iterative method described in subsection 4.2 is obtained.

5. Efficiency

In this section, sampling costs for a two-phase sample design using simple random sampling are compared to the costs of two one-phase alternatives. The one-phase designs considered are:

A - simple random sampling with stratification by size within each SIC2 x province cell;

B - simple random sampling with stratification using province and four digits of Revenue Canada SIC codes.

Subsection 5.1 includes definitions. The optimal allocation problems for designs A and B are formulated in subsection 5.2. The results of some comparisons of sampling costs are given in subsection 5.3.

5.1 Definitions

Consider one SIC2 *x* province cell containing *N* units. Each taxfiler in the population falls in one of *G* size strata. A SIC4 code is assigned to each taxfiler by Revenue Canada and each assigned code may or may not correspond to the true code. Let N_{ghk}, $h = 1,2,\ldots,H$, $k = 1,2,\ldots,K$, denote the number of units with true SIC4 code *h* and assigned SIC4 code *k* in size stratum *g*. Denote the total number of units in size stratum *g* with true SIC4 code *h* by $N_{gh.}$ and the total number of units in size stratum *g* with assigned SIC4 code k by $N_{g.k}$ and define $N_{h.} = \sum_g N_{gh.}$ and $N_{.k} = \sum_g N_{g.k}$. Let I_{ghk} denote the set of indices referring to taxfilers in size stratum *g* with true code *h* and assigned code *k* and define $I_{gh.} = I_{gh1} \cup I_{gh2} \cup \ldots \cup I_{ghK}$. Let Y_{hk} denote the population total of *y* over units with true SIC4 *h* and assigned SIC4 *k* and note that $Y(h) = \sum_k Y_{hk}$.

For each sample design, it is necessary to determine sample sizes that minimize costs while satisfying the constraints

$$V(\hat{Y}_{H\text{-}T}(h))/Y_h^2 \le C_h^2, \quad \forall h,$$

where $\hat{Y}_{H\text{-}T}(h)$ is the Horvitz-Thompson estimator of the population total of *y* for SIC4 domain *h*. Sampling costs, K_1 and K_2, do not vary between strata.

5.2 Optimal Allocation

The Horvitz-Thompson estimator for design A is

$$\hat{Y}_A(h) = \frac{N}{n} \cdot \sum_g \sum_{i \in s \cap I_{gh.}} y_i,$$

where $\sum_{i \in s \cap I_{gh.}}$ denotes the sum over all sampled units in size group *g* and SIC4 domain *h*. The variance of this estimator is

$$V(\hat{Y}_A(h)) = \sum_g \frac{N_g^2}{n_g} \cdot (1 - n_g/N_g) \cdot S^2_{gh.},$$

where

$$S^2_{gh.} = \frac{1}{(N_g - 1)} \cdot \left(\sum_{i \in I_{gh.}} y_i^2 - \frac{Y^2_{gh.}}{N_g} \right).$$

Using the transformation of variables $X_g = 1/\upsilon_g - 1$, where $\upsilon_g = n_g/N_g$, the optimal allocation problem for design A can be written as a convex programming problem involving the minimization of

$$F_A = \sum_g \frac{N_g}{X_g + 1} \cdot (K_1 + K_2),$$

subject to the constraints

$$C_h^2 \cdot Y_h^2 - \sum_g X_g \cdot (A_{gh} + B_{gh}) \ge 0, \quad \forall h,$$

$$X_g \ge 0, \quad \forall g$$

where A_{gh} and B_{gh} are given by (3).

Design B involves selection of $n_{g.k}$ units from the $N_{g.k}$ units that are in size group *g* and have been coded to SIC4 *k* by Revenue Canada. The Horvitz-Thompson estimator is a stratified domain estimator given by

$$\hat{Y}_B(h) = \sum_g \sum_{k=1}^{K} \frac{N_{g.k}}{n_{g.k}} \sum_{i \in s \cap I_{ghk}} y_i.$$

The variance of this estimator is

$$V(\hat{Y}_B(h)) = \sum_g \sum_{k=1}^{K} \left(\frac{N_{g.k}}{n_{g.k}} - 1\right) \cdot N_{g.k} \cdot S^2_{ghk},$$

where

$$S^2_{ghk} = \frac{1}{(N_{g.k} - 1)} \cdot \left(\sum_{i \in I_{ghk}} y_i^2 - \frac{Y^2_{ghk}}{N_{g.k}} \right).$$

Using the transformation $X_{g.k} = 1/\upsilon_{g.k} - 1$, where $\upsilon_{g.k} = n_{g.k}/N_{g.k}$, the optimal allocation problem for design B can be formulated as a convex programming problem. It is necessary to minimize the cost function

$$F_B = \sum_g \sum_{k=1}^{K} \frac{N_{g.k}}{X_{g.k} + 1} \cdot (K_1 + K_2),$$

with respect to $X_{g.k}$, $g = 1,2,\ldots G$, $k = 1,2,\ldots,K$, subject to the constraints

$$C_h^2 \cdot Y^2_{h.} - \sum_g \sum_{k=1}^{K} X_{g.k} \cdot N_{g.k} \cdot S^2_{ghk} \ge 0, \quad \forall h,$$

$$X_{g.k} \ge 0, \quad \forall k.$$

The optimal allocation problem for the two-phase design has been described in section 4.

5.3 Comparison of Methods

To compare the methods, data for the province of Québec for TY(1988) was used without size stratification (G=1). It was assumed that SIC4 codes assigned by Statistics Canada were 100% accurate. When population totals and variances for SIC4 domains were required by the allocation methods, estimates based on the TY(1988) first-phase sample were employed.

In the case $G=1$, the sample size required using design A is

$$\tilde{n}_A = \max_h \{N^2 \cdot S_h^2 / (C_h^2 \cdot Y_h^2 + N \cdot S_h^2)\}.$$

(Note that the subscript g has been dropped.) The corresponding minimum sampling cost is

$$K_A = (K_1 + K_2) \cdot \tilde{n}_A.$$

Let $\tilde{n}_{Bk}$ denote the optimal sample size in stratum k for design B in the case $G=1$. The computer code for the method of Schittkowski (1985) available in IMSL (1987) was used to solve the convex programming problem required to obtain optimal sample sizes. The minimum cost for design B is

$$K_B = (K_1 + K_2) \cdot \sum_{k=1}^{K} \tilde{n}_{Bk}.$$

In the case $G=1$, the minimum cost for the two-phase design is

$$K_{2\text{-}PH} = K_1 \cdot \tilde{n} + K_2 \cdot \sum_{h=1}^{H} \tilde{n}_h,$$

where

$$\tilde{n} = \left(\max_h \left\{ \frac{A_h + B_h}{C_h^2 \cdot Y_h^2 + A_h + B_h} \right\} \right) \cdot N,$$

and

$$\tilde{n}_h = \frac{A_h \cdot \tilde{n} \cdot N_h}{(\tilde{n} - N) \cdot B_h + \tilde{n} \cdot A_h + \tilde{n} \cdot C_h^2 \cdot Y_h^2}, \quad \forall h.$$

Information about the ratio of first- and second-phase sampling costs is needed. Tax return information is microfilmed or photocopied at Revenue Canada. SIC4 codes are assigned to businesses at Statistics Canada. The cost of microfilming or photocopying is approximately $2.80 per taxfiler and the cost of SIC4 coding is $0.95 per taxfiler. The cost of capturing financial data for a taxfiler in the second-phase sample is $3.75. SIC4 codes assigned to businesses and used for second-phase sample allocation cannot be updated in subsequent years using information from the second-phase sample. An review mechanism that is independent of second-phase sample selection is required.

The cost of selecting a taxfiler for the first time for the first-phase sample or reviewing an SIC4 code is $3.75. The second-phase sampling cost is also $3.75 for taxfilers selected for the first time for the first-phase sample as well as taxfilers being reviewed. It is $7.50 for taxfilers that were selected for the first-phase sample for the first time in an earlier tax year and are not being reviewed in the current year. A ratio of 1:1 for first- to second-phase sampling costs would be appropriate if SIC4 codes for all businesses in the first-phase sample were reviewed annually. If codes were reviewed every third year and there were no businesses that ceased operations, the appropriate ratio would be 1:5. An appropriate regime for review of SIC4 codes depends on the deterioration in the quality of codes over time. A review scheme has not yet been implemented.

Costs for design B and the two-phase approach, relative to design A, are given in Table 1. The rightmost column of the table gives relative cost figures that would be obtained if all SIC codes assigned by Revenue Canada were accurate to four digits. The cost of design B is four times greater than the cost would be in the absence of SIC coding error. The two-phase design is comparable to design B for a 1:1 ratio of first- and second-phase sampling costs and is more efficient for a 1:5 ratio.

For a given ratio of sampling costs, the cost of the two-phase approach relative to design B depends on the frequency of errors in the third and fourth digits of SIC codes assigned by Revenue Canada. The error rate for the Québec data was 28%. The effects on relative costs of various types of errors and different error rates is a subject for future work.

K_1:K_2	Sample Design		
	Two-Phase	B	B (No errors)
1:1	51%	48%	12%
1:5	26%	48%	12%

Table 1: Sampling costs relative to design A

6. Conclusion

The results in this paper suggest that the efficiency of the two-phase sample design compares favourably with alternatives. To conclude on a cautionary note it is necessary to mention that the practical effectiveness of the allocation methodology described here, as well as the validity of the efficiency comparison, depends on the availability of accurate SIC codes. Accuracy is needed in the first two digits of SIC codes supplied by Revenue Canada as well as the SIC4 codes assigned to businesses in the first-phase sample by Statistics Canada and used for second-phase sample selection.

References

ARMSTRONG, J., BLOCK, C. and SRINATH, K.P. (1991). Two-phase sampling of tax records for business surveys. Statistics Canada, Methodology Branch Working Paper (forthcoming).

ARMSTRONG, J. and LE PETIT, C. (1991). Répartition de l'échantillon pour un plan de sondage à deux phases, avec application aux données fiscales. Statistics Canada, Methodology Branch Working Paper BSMD-91-001F.

CHOUDHRY, G.H., LAVALLEE, P., and HIDIROGLOU, M. (1989). Two-phase sample design for tax data. American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 646-651.

COCHRAN, W.G. (1977) *Sampling Techniques*, (3rd Ed.). New York: John Wiley.

ESTEVAO, V., and TREMBLAY, J. (1986). An evaluation of the assignment of standard industrial codes from T2 tax data. Business Survey Redesign Project Working Paper, Statistics Canada.

IMSL (1987). Math/Library FORTRAN subroutines for mathematical applications. Houston: IMSL Inc.

SCHITTKOWSKI, K. (1985). NLPQL: a FORTRAN subroutine solving constrained nonlinear programming problems. *Annals of Operations Research*, 5, 485-500.

SMITH, P.J. (1989). Is two-phase sampling really better for estimating age composition?. *Journal of the American Statistical Association*, 84, 916-921.

SUNTER, A.B. (1986). Implicit longitudinal sampling from administrative files: a useful technique. *Journal of Official Statistics*, 2, 161-168.

WU, J. (1989). Comments reported in Minutes of the Meeting of the Advisory Committee on Statistical Methods, April 3-4, 1989. Unpublished document, Statistics Canada.