Survey Methodology, December 1991
Vol. 17, No. 2, pp. 195–210
Statistics Canada

A Sampling and Estimation Methodology for Sub-Annual Business Surveys

M.A. HIDIROGLOU, G.H. CHOUDHRY and P. LAVALLÉE[1]

ABSTRACT

A sample design for the initial selection, sample rotation and updating for sub-annual business surveys is proposed. The sample design is a stratified clustered design, with the stratification being carried out on the basis of industry, geography and size. Sample rotation of the sample units is carried out under time-in and time-out constraints. Updating is with respect to the selection of births (new businesses), removal of deaths (defunct businesses) and implementation of changes in the classification variables used for stratification, *i.e.* industry, geography and size. A number of alternate estimators, including the simple expansion estimator and Mickey's (1959) unbiased ratio-type estimator have been evaluated for this design in an empirical study under various survey conditions. The problem of variance estimation has also been considered using the Taylor linearization method and the jackknife technique.

KEY WORDS: Continuous surveys; Sample updating; Ratio estimator; Variance estimation.

1. INTRODUCTION

The universe for sub-annual business surveys continually changes on account of births, deaths, splits, mergers, amalgamations, and classification changes. The sample design associated with such a universe should have the following characteristics. Firstly, it should result in samples which reflect the changing structure of the population. Secondly, it should distribute response burden by rotating units in and out of the sample. Thirdly, if there are significant changes in the stratification of the universe, it should be possible to redraw a new sample which reflects the stratification and possible changes in sampling fractions. The resulting new sample should have maximum overlap with the previous sample in order to minimize abrupt changes in the estimates and increased costs due to the introduction of new units in the sample. The sample design which has been proposed to satisfy these requirements is that of a simple random sample of randomly formed rotation groups (clusters) within each of the strata. Each rotation group represents either a group of units or a single unit. All units within a selected rotation group are selected in the sample. Rotation of the sample takes place under the constraints that units must stay in the sample for a certain period of time and be kept out of the sample for at least a certain period of time after they have rotated out of the sample.

For given domains of interest, unbiased (or nearly unbiased) estimates are developed along with the associated measures of reliability (coefficients of variation). A desirable property of the estimation is that the estimates of domain totals should add up to the population total when the domains are exhaustive and non-overlapping. This can be ensured by using one set of weights which is independent of the domains.

In section 2, the rotation group sampling design is developed and a number of alternative estimation procedures are described in section 3. In section 4, the results of an empirical study showing the performance of these estimators under various survey conditions are given. Finally, section 5 contains some concluding remarks.

[1] M.A. Hidiroglou, G.H. Choudhry and P. Lavallée, Business Survey Methods Division, Statistics Canada, 11th floor R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6.

2. SAMPLING DESIGN

2.1 Stratification and Sample Allocation

The stratification of a business universe is usually based on one or more of the following characteristics: industry, geography, and size. The size measure can be univariate (*e.g.* sales or number of employees) or multivariate (*e.g.* revenue and assets). In our context, the primary strata are cross-classifications of industry and geographic regions for which estimates are required. Within these primary strata, secondary strata are formed using the size measure of the units. The secondary strata are comprised of a completely enumerated "take-all" stratum and a number of strata called 'take-some' strata where sampling occurs. It is necessary to have a take-all stratum on account of the highly skewed nature of the business universe. The take-all stratum boundary can be determined by a method introduced by Hidiroglou (1986). This method finds the optimum boundary between the take-all and the take-some strata within each primary stratum so as to minimize the overall sample size for a given coefficient of variation. The determination of this boundary also takes into account that certain units are to be sampled with certainty irrespective of their size. These pre-specified "take-all" units are units which are to be included in the sample on account of their complex structures. An example of a unit with a complex structure could be one which operates in more than one of the primary strata. The boundaries for the take-some strata are obtained either using the cum $\sqrt{f}$ rule introduced by Dalenius and Hodges (1959) or the cum $\sqrt{x}$ rule given by Hansen *et al.* (1953). Here x is a size variable available for stratification of the units in the population.

The sample sizes for the primary strata are computed so as to satisfy planned levels of precision for certain key estimates. The computation of these sample sizes also takes into account the required allocation scheme of the units to the take-some strata. It is assumed that the information available for computing these sample sizes is well correlated with the planned key variables. Given that the take-all sample units have been taken into account, the remaining sample is allocated to the take-some strata within the primary stratum, proportional to M^q or X^q, where M is the number of units in the take-some stratum and X is the take-some stratum total for the size variable being considered. The power q where $0 \leq q \leq 1$ is chosen according to the required allocation. Letting $q = 1$ results in Neyman allocation, whereas as q approaches zero, the resulting coefficients of variation become more equal amongst the different strata provided that $S_h/\bar{X}_h$ does not vary significantly from stratum to stratum and that the finite population correction factors can be ignored. The advantages of these power allocations are discussed in Bankier (1988). The allocation can be adjusted to achieve the desired minimum sample sizes and/or maximum weights for each secondary stratum.

The reliability criteria (in terms of coefficients of variation) can be associated with the primary strata in one of two ways. Either they can be specified for each primary stratum, or, for a given global (national) coefficient of variation (c.v.), the c.v. at the primary stratum level can be determined so that the c.v.'s for each industry group and geographic region are equal. An iterative procedure is used to determine the desired c.v.'s for each of the primary strata and hence the sample size within each primary stratum, so that the planned c.v.'s at the global and marginal levels are achieved.

2.2 Sampling Scheme

For each stratum, the M population units within that stratum are randomly allocated to a predetermined number P of population rotation groups, so that initially, the number of units in each of any two rotation groups differ by at most one unit. The number of rotation groups is a function of sampling fractions, and time-in and time-out constraints. It may be noted that in order to achieve unbiasedness, the time-in and time-out constraints may sometimes have to be violated. A simple random (SRS) sample of p rotation groups is selected from the P population rotation groups. The number of rotation groups p to be selected is determined such that p/P is approximately equal to the desired sampling fraction f. The sample consists of all the units in the p selected rotation groups. Rotation of the sample occurs by acquiring an out-of-sample rotation group and dropping an in-sample rotation group. Births are randomly allocated to the P population rotation groups, one at a time, in a systematic fashion. Deaths are removed from the stratum only if they are detected by a source independent of the survey, or if they have been dead for more than a pre-specified period of time. Methods proposed by Kish and Scott (1971) are adopted for sample updating with re-stratification due to population units changing strata. The sample update maximizes the overlap between the current and the new samples. There are obvious advantages to redrawing the sample in this fashion. First, it minimizes the introduction of new units into the sample, resulting in a smoother transition from an operational point of view, and also minimizes cost. Second, discontinuity in the estimates on account of sample redraw is kept to a minimum.

There are other sampling schemes which can be used to select the sample and rotate the units. These include Poisson and collocated sampling. The properties of these schemes have been discussed by Brewer, Early and Joyce (1972), and by Sunter (1977). Poisson sampling as defined by Hajék (1964) allows each unit in the population to be drawn in the sample independently with a given probability of inclusion. Decisions as to whether the unit is selected in the sample or not are made using an independent random draw or Bernoulli trial for each unit. Supposing that the inclusion probability of a given unit i is π_i, and that a random number u_i uniformly distributed over the interval (0,1) is generated, then the i-th unit is selected if $u_i \leq \pi_i$. This probability of inclusion corresponds to the sampling fraction of the stratum that the unit belongs to. Although the advantage of Poisson sampling lies in the simple manner in which sample rotation is exercised, it has certain disadvantages. Its main disadvantage is that the realized sample size is a random variable. This can be serious if the number of units in the stratum is small, possibly resulting in samples of size zero. Early and Brewer (1971) remedied this weakness by using a scheme known as collocated sampling. Collocated sampling is similar to Poisson sampling but reduces the variation in sample size by equispacing, at the cell level, the units over the interval (0,1). Properties of this method are provided in more detail in Brewer, Early and Hanif (1984). Whereas in Poisson sampling, the addition of births and removal of deaths do not affect the random numbers attached to existing units, the use of collocated sampling requires that these random numbers be slightly perturbed, possibly disturbing the rotation scheme by violating the time-in and time-out constraints.

The rotation group sampling scheme has several advantages over the two previously mentioned schemes. For the rotation group sampling scheme, in contrast to the Poisson scheme, the expected number of units on each rotation cycle is almost equal. The removal of dead units on a universal basis may disturb the balance of units amongst the different rotation groups. This can be remedied by periodically redrawing the sample with maximum overlap, keeping the stratification and sampling fractions unchanged. The rotation for the rotation group scheme can be performed without perturbing the units, thereby satisfying the time-in and time-out constraints. This may not necessarily be true with collocated sampling on account of the slight perturbations of the random numbers due to population births and deaths. These effects may become non-trivial over a long period of time. Another advantage of the rotation group scheme over the other two methods is that re-stratification and new sampling fractions can easily be accommodated while maximizing sample overlap.

2.2.1 Determination of the Number of Rotation Groups

Assume that for a given take-some stratum, the number of population units is M and that the desired sampling fraction is f. Let t_{in} be the desired number of occasions a unit should stay in the sample. Let t_{out} be the minimum required number of occasions a unit must stay out of the sample, once it has rotated out of the sample. The required number of population rotation groups "P" and *in*-sample rotation groups "p" are determined as follows. Let, $x = int\,[t_{in}\,(1 - f)/f + 0.5]$ where $int[\cdot]$ denotes the integer portion of the argument. Two conditions arise:

a) If $x \geq t_{out}$, then the number of in-sample rotation groups is $p = t_{in}$ and the number of population rotation groups is $P = t_{in} + x$.

b) If $x < t_{out}$, then the number of in-sample rotation groups is

$$p = int\left[\frac{f}{1 - f}t_{out} + 0.5\right]$$

and the number of population rotation groups is $P = p + t_{out}$.

It must be noted that p/P is only approximately equal to f on account of the integer operations.

2.2.2 Allocation of Units to Rotation Groups

Given that at the time of initial selection, there are M population units to be allocated to P population rotation groups, two distinct cases arise with respect to the relative sizes of M and P: $M \geq P$ or $M < P$.

When $M \geq P$, at least one unit can be allocated to each population rotation group. Suppose $M = a\,P + \ell$, where $a > 0$ and $\ell \geq 0$ are integers. In order to equalize the rotation group sizes as much as possible at the time of initial selection and on subsequent occasions, the following procedure is used. A 2 by P matrix is used to assign a rotation sequence to the units that will satisfy the requirements of almost equal rotation group sizes. It is used for initial sample selection and subsequent addition of births. The first "assignment" row is labelled from 1 to P, whereas the second "rotation" row is a randomized order of the first row. The corresponding rotation group numbers in the second row determine which units are in sample at any point in time. The M population units are assigned sequentially to the assignment rotation group numbers $1, 2, \ldots, P$, the P-th unit going to the P-th assignment rotation group number. The $(P + 1)$-th unit is assigned to assignment rotation group number 1 and so on. This eventually results in having the first "ℓ" assignment rotation groups with $(a + 1)$ units and the next $(P - \ell)$ assignment rotation groups with (a) units. The rotation group to which the M-th unit is assigned is termed the last assignment rotation group. This rotation group, which is assigned rotation group number at time of initial selection, is used for assigning future births starting from the next assignment rotation group number, *i.e.* $\ell + 1$.

When $M < P$, the M population units can only be allocated to a subset of M out of the P rotation groups. These rotation groups must be as equispaced as possible to ensure that the expected sample size, $\bar{m} = fM$, will be achieved from one survey occasion to the next. For this case the allocation matrix is 2 by M. The first assignment row is labelled from 1 to M. The second rotation row is a randomization of M "z" numbers where $1 \leq z_i < z_j \leq P$ for $i \neq j$, $i = 1, \ldots, M$ and $j = 1, \ldots, M$. The "z" numbers are created as follows.

i) Find integers s and q such that $P = sM + q$ where $q < M$ and $s \geq 0$.

ii) Generate r_j $(j = 1, \ldots, M)$ numbers randomly assuming the values 0 or 1, such that q of them have the value equal to 1 and $M - q$ of them have the value equal to 0.

iii) Select a random integer "b" such that $1 \leq b \leq P$.

iv) Compute $z_1 = (b + r_1 - 1) \bmod P + 1$ and $z_j = (z_{j-1} + s + r_j - 1) \bmod P + 1$ for $j = 2, \ldots, M$.

v) Randomize the "z" numbers. Let the sequence of randomized "z" numbers be $z_{i_1}, z_{i_2}, \ldots, z_{i_M}$.

Now the M population units are assigned sequentially to the M assignment rotation group numbers, thereby picking up their rotation numbers. The last assignment rotation group number is M. Future births will be assigned starting from assignment rotation group number 1.

It is now a simple matter to perform the basic functions of sample selection and updating.

2.2.3 Sample Selection and Updating

At time of initial sample selection, a given stratum will have $N = \min(M,P)$ distinct rotation groups. The units belonging to the initial sample are those whose rotation numbers are included in the closed sampling interval $[1,p]$. When $M \geq P$, the number of in-sample rotation groups n is equal to p. When $M < P$, the number of in-sample rotation groups n is approximately equal to fN on account of the equispacing.

Sample rotation is carried out by shifting the sampling interval by one rotation group at each sampling occasion in a circular fashion. On the t-th occasion, units in the sample are those whose rotation number is contained in the interval defined as

i) $[(t - 1) \bmod P + 1,\ (t + p - 2) \bmod P + 1]$, if $(t - 1) \bmod P \leq (P - p)$

and

ii) $[1, (p - P) + (t - 1) \bmod P] \cup [(t - 1) \bmod P + 1, P]$, otherwise.

Effectively, rotation occurs by dropping a rotation group from in-sample and acquiring a rotation group from out-of-sample in a modular fashion.

"Births" occur as a result of starting a new business activity, or a change of industrial activity of a unit from out-of-scope to in-scope for the survey. Births are stratified and given an assignment rotation group number within the stratum as follows. Assuming the last assignment rotation group number was ℓ, where $1 \leq \ell \leq P$, the q-th birth will be given the assignment rotation group number $(\ell + q) \bmod P$. The next birth will be given the assignment rotation group number $(\ell + q + 1) \bmod P$. The rotation number is then immediately obtained through the one-to-one correspondence between the assignment and rotation numbers.

"Deaths" occur as a result of the termination of business activity for in-scope units or changes of industrial activity from in-scope to out-of-scope to the survey. Deaths that occur in a take-all stratum are immediately removed from the population and sample. Deaths that are part of a take-some stratum are removed immediately if they are identified as such by a source independent of the survey process. Otherwise, they are removed after a given time period. This time period should be sufficiently long so that most of the population deaths would have been identified. Deaths in the sample and in this latter category which have not yet been removed are assigned a value of zero for estimation purposes. Classification values are also retained as such until they have been identified as changes by a source independent of the survey.

2.2.4 Periodic Resampling

The sampling frame changes continually due not only to births and deaths, but also due to changes of classification variables used in the stratification (*i.e.*, geography, industry and size). These changes in the classification variables are reflected in the estimation process by use of domain estimation (*i.e.* estimation for sub-populations). That is, the latest classification is assigned to data for tabulation purposes, using the original sampling weight. Over a period of time, changes in classification may be sufficiently important to require the examination of the stratification and subsequent sampling rates. One solution would be to redraw an independent sample, taking into account these changes, but ignoring the current sample. Such an approach has certain disadvantages from an operational point of view. An independent redraw implies that i) the newly sampled units must be initiated into the sample, ii) time-in and time-out constraints can be violated, and iii) the estimates may change substantially. It is therefore desirable to maximize the overlap between the current sample and the new sample. The following methodology provides such a procedure for resampling. It is an adaptation of the Kish and Scott (1971) method, and is based on the property that each rotation group is a simple random sample from the population rotation groups.

At time of resampling, rotation will have occurred at different rates amongst the strata, resulting in sampling intervals with different starting and end points. Hence, assuming that rotation started at time t_1 and that we are currently at time t_2, the number of rotations that have occurred is $r = t_2 - t_1 + 1$. At time t_2, the sampling interval(s) associated with a given stratum currently labelled as k ($k = 1, 2, \ldots, K$) is(are)

$$[(r-1)\bmod P_k + 1,\ (r + p_k - 2)\bmod P_k + 1] \quad \text{if} \quad (r-1)\bmod P_k \le (P_k - p_k)$$

and

$$[1, p_k - P_k + (r-1)\bmod P_k] \quad \text{and} \quad [(r-1)\bmod P_k + 1, P_k] \quad \text{otherwise.}$$

The first step associated with the resampling is to relabel the different sampling intervals, which have different starting points, into sampling intervals which have the same starting point. For the k-th stratum, the resulting sampling interval is $[1, p_k]$. Let b denote the starting point of the sampling interval at time t_2 where b is given by $(r - 1)\bmod P_k + 1$. All units labelled with rotation number "g" are relabelled as $(g - b + 1)$ if $b \le g \le P_k$ and as $P_k - (b - g - 1)$ otherwise.

The second step is to associate with each population unit currently classified to stratum k its new stratum "h". The population units of the new h-th stratum, U_h, can therefore be expressed as the union of K non-overlapping and exhaustive sets U_{hk}, $h = 1, 2, \ldots, L$. Each set U_{hk} is comprised of population units whose new stratification is h and current stratification is k. Some of these sets may be empty.

The third step is to rank, on the 0 to 1 scale, sampling units within each set U_{hk}, taking into account their current rotation numbers. Assume that there are M_{hk} units in the set U_{hk} and that their current rotation numbers are labelled between 1 and P_k. Rank these units from 1 to M_{hk} based on their associated current rotation number. Units which have the lowest rotation numbers are assigned the lowest ranks and units which have the highest rotation numbers are assigned the highest ranks. If there are any ties, these can be broken up randomly by generating uniform random numbers. This results in the units in set U_{hk} to be ranked from 1 to M_{hk}. Next, a unit with rank "i" in set U_{hk}, $1 \le i \le M_{hk}$, is assigned a number $r_{hki} = (a_{hk} + i - 1)/M_{hk}$, where a_{hk} is a uniformly generated random number between 0 and 1 for each set U_{hk} within U_h. These numbers represent the current rotation groups transformed to the range 0 and 1. Assume that the new sampling fraction associated with the new stratum is f_h and that the current sampling fraction is f_k. If $f_h \ge f_k$, this implies that all units currently sampled in U_{hk} will stay in the new sample and that units in the closed interval $[0, f_h]$ will be included in the new sample. If $f_h < f_k$, this implies that units must be dropped (rotated out) from the current sample. The units which must be dropped are those which have the lowest r_{hki} values. These represent the rotation groups which have been in the sample the longest. In order that the units in the new sample be contained in the closed interval $[0, f_h]$ it is necessary to relabel the r_{hki}'s as $r_{hki} - (f_k - f_h)$ if $r_{hki} \geq (f_k - f_h)$ and as $r_{hki} - (f_k - f_h) + 1$ otherwise. Assuming that the population units belonging to the new h-th stratum are ranked based on the ordered r_{hki}'s, define $b_{hi} = i/(M_h + 1), i = 1, 2, \ldots, M_h$. Using the b_{hi}'s, new rotation numbers will be obtained as follows. For a given new stratum h, let N_h be the number of distinct rotation groups. Form N_h disjoint intervals

$$I_u = \begin{cases} [(u-1)/N_h, u/N_h] & \text{for} \quad u=1, \ldots, N_{h-1} \\ [(N_h - 1)/N_h, 1] & \text{for} \quad u = N_h. \end{cases}$$

The union of these intervals is the closed interval $[0,1]$ D_{u_i}. For D_{u_j} the new stratum h, label the new rotation numbers as where $D_1, D_2, \ldots, D_{N_h}$ where $D_{u_i} < D_{u_j}$ for $u_i < u_j$, $u_i = 1, \ldots, N_h$. The i-th unit acquires rotation number D_u if its corresponding b_{hi} value belongs to the interval I_u. Assuming that all the M_h units have been assigned new rotation numbers in this fashion, the units in sample will be those whose rotation number belongs to the interval $[1, p_h]$.

3. WEIGHTING AND ESTIMATION

The simplest estimator which can be used in conjunction with the rotation group design described in Section 2.2 is the simple expansion (or simple domain) estimator. Although this estimator is unconditionally unbiased, it can have a large conditional bias when the rotation group sizes are not balanced. The removal of dead units can cause such an imbalance in the distribution of rotation group sizes. Other estimators which take the auxiliary rotation group size information into account have therefore been considered. These include the separate and combined ratio estimators. A drawback of the separate estimator is that its bias may accumulate in a non-trivial manner across strata. The combined estimator will have negligible bias, but possibly large variances for stratum level estimates. We have therefore evaluated the performance of an unbiased separate ratio estimator due to Mickey (1959). The penalty for achieving unbiasedness is an increase in the variance. The primary objective is to determine which of the above estimators is the most suitable one for the rotation group design. The criteria for choosing the most appropriate estimator will be based on bias and mean squared error. In order to simplify the comparisons, it will be assumed that each sampled unit has valid response data.

As mentioned earlier, the h-th stratum ($h = 1, 2, \ldots, L$) is defined at some given level of industry, geography and size. Estimates are required for domains which can span all the sampling strata or be a subset of these strata. Examples of such domains are aggregations of variables of interest at the sub-provincial level given that the sampling may have occurred at a higher level, *e.g.* province. A desirable feature of the estimates is that the sum of any non-overlapping domain set must always add up to the domain defined as their union. In order to achieve consistency, only one set of weights can be used.

Let y denote the characteristic of interest and y_{hij} be its value for the j-th unit in rotation group (cluster) i of stratum h. Let $\delta_{hij}(d)$ be an indicator variable defined as 1 if the hij-th unit belongs to domain "d", and 0 otherwise. Then, the parameter of interest is the population total $Y(d)$ given by:

$$Y(d) = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} y_{hij}(d),$$

where $y_{hij}(d) = \delta_{hij}(d)\, y_{hij}$.

As described earlier, we have a simple random sample of n_h rotation groups selected without replacement from the N_h rotation groups in the h-th stratum. Let M_{hi} be the number of units in the i-th sampled rotation group within stratum h. Without loss of generality, we can assume that the sampled rotation groups are indexed $i = 1, 2, \ldots, n_h$. Let $y_{hi}(d)$ be the total response of the units belonging to domain "d" from the i-th sampled rotation group within stratum h, *i.e.*

$$y_{hi}(d) = \sum_{j=1}^{M_{hi}} y_{hij}(d),\ i = 1, 2, \ldots, n_h.$$

We will consider a number of alternative estimators for the population parameter $Y(d)$ and their corresponding variance. The estimators considered are of the form,

$$\hat{Y}_h(d) = \sum_{i=1}^{n_h} w_{hi}\, y_{hi}(d),$$

where w_{hi} is the product of the design weight and an adjustment which reflects the estimation procedure used. Estimators of $Y(d)$ are obtained by aggregating over strata, that is,

$$\hat{Y}(d) = \sum_{h=1}^{L} \hat{Y}_h(d).$$

3.1 Estimators of Total

A. Simple Expansion Estimator

Since the probability of selecting a rotation group in the h-th stratum is n_h/N_h, the design weight is $w_{hi} = N_h/n_h$ for $i = 1, 2, \ldots, n_h$, $h = 1, 2, \ldots, L$. The simple expansion estimator is given by

$$\hat{Y}_E(d) = \sum_{h=1}^{L} N_h\, \bar{y}_h(d), \tag{3.1}$$

where

$$\bar{y}_h(d) = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}(d).$$

As mentioned earlier, this estimator is unconditionally unbiased, but it can have a large conditional bias. Moreover, it may not be very efficient because it does not make use of available auxiliary information, such as rotation group sizes. As the variation in the rotation group sizes may increase over time on account of removal of deaths, it may become more and more inefficient.

B. Separate Ratio Estimator

If the correlation between $y_{hi}(d)$ and rotation group sizes M_{hi} is large, efficiency gains can be realized through the separate ratio estimator defined as

$$\hat{Y}_{SR}(d) = \sum_{h=1}^{L} \left(\frac{M_h}{\bar{m}_h}\right) \bar{y}_h(d) \tag{3.2}$$

where

$$\bar{m}_h = n_h^{-1} \sum_{i=1}^{n_h} M_{hi}$$

and

$$M_h = \sum_{i=1}^{N_h} M_{hi}.$$

One major drawback of this estimator is that it is subject to the ratio estimation bias. Consequently, if the bias tends to be positive or negative in the majority of the strata, its accumulated effect can be quite significant when aggregating over the strata.

C. Combined Ratio Estimator

The accumulated effect of aggregation bias can be significantly reduced using a combined version of the ratio estimator. The combined ratio estimator is given by

$$\hat{Y}_{CR}(d) = M \frac{\sum_{h=1}^{L} N_h \bar{y}_h(d)}{\sum_{h=1}^{L} N_h \bar{m}_h}, \tag{3.3}$$

where $M = \sum_{h=1}^{L} M_h$.

D. Unbiased Ratio-type Estimator

The bias problem caused by the ratio estimation can be completely eliminated using the following adjusted ratio-type estimator suggested by Mickey (1959). The Mickey estimator is given by

$$\hat{Y}_{MI}(d) = \sum_{h=1}^{L} \left(\bar{r}_h(d) M_h + (N_h - n_h + 1) \left[\sum_{i=1}^{n_h} y_{hi}(d) - m_h \bar{r}_h(d) \right] \right), \tag{3.4}$$

where

$$\bar{r}_h(d) = \frac{1}{n_h} \sum_{j=1}^{n_h} \bar{r}_h^{(j)}(d); \; \bar{r}_h^{(j)}(d) = \frac{\sum_{i \neq (j)} y_{hi}(d)}{\sum_{i \neq (j)} M_{hi}}; \; m_h = \sum_{i=1}^{n_h} M_{hi}.$$

An undesirable feature of the Mickey estimator is that it can have weights less than one, including negative weights.

For the separate and combined ratio estimators, the variances are estimated using the Taylor linearization method. In the case of Mickey's estimator, a jackknife procedure is used, leaving out one rotation group at a time and re-computing Mickey's estimator for the remaining $(n_h - 1)$ rotation groups in the sample. Denote each jackknifed estimator as for $\hat{Y}_{MI,h}^{(j)}(d)$ for $j = 1, 2, \ldots, n_h$,

where

$$\hat{Y}_{MI,h}^{(j)}(d) \sum_{i \neq (j)} w_{hi}^{(j)} y_{hi}(d)$$

with

$$w_{hi}^{(j)} = [M_h - (m_h - M_{hj})] \; (N_h - n_h + 2) \; b_{hi}^{(j)} + (N_h - n_h + 2)$$

and

$$b_{hi}^{(j)} = (n_h - 1)^{-1} \sum_{i \neq (j)} \frac{1}{(m_h - M_{hj} - M_{hi})}.$$

A jackknife variance estimator of $\hat{Y}_{MI,h}(d)$ is given by

$$v_j(\hat{Y}_{MI,h}(d)) = (1 - f_h) \frac{(n_h - 1)}{n_h} \sum_{j=1}^{n_h} (z_h^{(j)}(d) - \bar{z}_h(d))^2,$$

where $z_h^{(j)}(d) = \hat{Y}_{MI,h}^{(j)}(d)$ and $\bar{z}_h(d) = n_h^{-1} \sum_{j=1}^{n_h} z_h^{(j)}(d)$.

It can be shown that all the estimators are equivalent and unconditionally unbiased when the rotation group sizes M_{hi} are all equal in each stratum h. However once the rotation group sizes (M_{hi}) become unequal, all estimators, except for the simple expansion and the Mickey estimator, are unconditionally biased. For these estimators, the magnitude of their unconditional biases and their efficiency was assessed in a simulation study which is presented next.

4. SIMULATION STUDY

The purpose of this simulation was to determine which of the four estimators of aggregate total $Y(d)$ and the stratum total $Y_h(d)$ would be the most "appropriate" for the sample design described in Section 2. For simplicity, the simulations were confined to a single variable (y), gross business income (GBI). Also, for the purpose of this simulation the domains coincided with strata. Therefore, the symbol "d" used to denote the domain will be omitted.

4.1 Description of the Study

The universe for the simulation study was defined as the set of smaller sized units belonging to the Wholesale Trade sector in the province of Québec for the May 1989 reference period. The size of each unit was based on a GBI derived from payroll deductions using a ratio model. Units whose GBI was below a given threshold were retained, resulting in a population of 10,953 units. The stratification of this population was defined on the basis of Standard Industrial Classification at the 3 digit level. This resulted in 30 strata with a minimum stratum size of 18 units. For each of the 30 strata, 16 rotation groups were formed by randomly assigning the units to the rotation groups as described in Section 2.2.2.

For each stratum h, samples of 4 rotation groups were obtained from the 16 rotation groups using simple random sampling without replacement. From each stratum there were 1,820 possible samples of size 4. Over the 30 strata there were 54,600 (30 strata times 1,820 samples per stratum) possible different estimates for the separate ratio estimation procedure. On the other hand, for the combined ratio estimator, a total of $(1{,}820)^{30}$ different estimates could be produced. For the simple expansion, the separate ratio estimator, and Mickey's estimator, all 54,600 possible samples were drawn. For the combined ratio estimator, 100,000 samples were randomly drawn from the $(1{,}820)^{30}$ possible samples.

4.2 Evaluation Criteria

The evaluation criteria involved bias and mean squared error. These are described next.

For each selected sample k, an estimate $\hat{Y}_h^{(k)}$ was produced for each stratum h and for each of the four estimators. The stratum expectation $E(\hat{Y}_h^{(k)})$ of this estimate was obtained as

$$E(\hat{Y}_h) = \frac{1}{K} \sum_{k=1}^{K} \hat{Y}_h^{(k)},$$

where K is the total number of samples drawn. It should be noted that for estimators (3.1) - (3.2) and (3.4), $E(\hat{Y}_h)$ was in fact the true expectation since all possible samples were drawn. For the combined ratio estimator (3.3), it corresponded to an unbiased estimate of the expectation. The resulting stratum bias was

$$\text{Bias}(\hat{Y}_h) \doteq E(\hat{Y}_h) - Y_h.$$

The total bias, Bias$(\hat{Y})$, was obtained by summing the stratum bias over all strata.

For estimators (3.1) - (3.2) and (3.4), we have that

$$\text{Var}(\hat{Y}_h) \doteq \frac{1}{K} \sum_{k=1}^{K} (\hat{Y}_h^{(k)} - E(\hat{Y}_h))^2$$

and

$$\text{Var}(\hat{Y}) = \sum_{h=1}^{L} \text{Var}(\hat{Y}_h).$$

For the combined ratio estimator (3.3), we have that

$$\text{Var}(\hat{Y}_h \doteq \frac{1}{K-1} \sum_{k=1}^{K} (\hat{Y}_h^{(k)} - E(\hat{Y}_h))^2$$

and

$$\text{Var}(\hat{Y}) = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{Y}^{(k)} - E(\hat{Y}))^2,$$

where $\hat{Y}^{(k)} = \sum_{h=1}^{L} \hat{Y}_h^{(k)}$ and $E(\hat{Y}) = \sum_{h=1}^{L} (\hat{Y}_h)$.

Finally, the stratum mean squared error, MSE$(\hat{Y}_h)$, of each estimator was defined as

$$\text{MSE}(\hat{Y}_h) = \text{Var}(\hat{Y}_h) + (\text{Bias}(\hat{Y}_h))^2$$

while the aggregate mean squared error, MSE$(\hat{Y})$, of each estimator was given by

$$\text{MSE}(\hat{Y}) = \text{Var}(\hat{Y}) + (\text{Bias}(\hat{Y}))^2.$$

Four criteria were used in comparing the relative behaviour of the proposed estimators. The first criterion was absolute relative bias. The stratum average absolute relative bias was computed as

$$\overline{\text{ARB}} = \frac{1}{L} \sum_{h=1}^{L} \left| \text{Bias}(\hat{Y}_h) \right| \Big/ Y_h,$$

while the aggregate absolute relative bias was computed as

$$\text{ARB} = \left| \sum_{h=1}^{L} \text{Bias}(\hat{Y}_h) \right| \Big/ Y,$$

where

$$Y = \sum_{h=1}^{L} Y_h .$$

The second criterion was the ratio of absolute bias to standard error which was called "absolute standard bias". The stratum average absolute standard bias was computed as

$$\overline{\text{ASB}} = \frac{1}{L} \sum_{h=1}^{L} \left| \text{Bias}(\hat{Y}_h) \right| \Big/ \sqrt{\text{Var}(\hat{Y}_h)}$$

while at the aggregate level, it was computed as

$$\text{ASB} = | \text{ Bias}(\hat{Y}) \text{ } | \Big/ \sqrt{\text{Var}(\hat{Y})} \, .$$

Following Cochran (1977), a reasonable value for the maximum acceptable bias over the standard error should not exceed 10% . Indeed, since the precision of an estimator is usually measured by its variance and not by its MSE, too large a bias as compared to the standard deviation would give a false impression of the precision of the estimator used.

The third criterion was efficiency, defined as the ratio of the root mean squared error of the estimator under study, RMSE($\hat{Y}^{EST}$), to that of the simple expansion estimator RMSE($\hat{Y}^{EST}$). The stratum average relative efficiency was computed as

$$\overline{\text{EFF}} = \frac{1}{L} \sum_{h=1}^{L} \left\{ \text{RMSE}\left(\hat{Y}_h^{EXP}\right) \Big/ \text{RMSE}\left(\hat{Y}_h^{EST}\right) \right\},$$

while at the aggregate level, the relative efficiency was computed as

$$\text{EFF} = \text{RMSE}\left(\hat{Y}^{EXP}\right) \Big/ \text{RMSE}\left(\hat{Y}^{EST}\right) .$$

Finally, the fourth criterion was to observe the proportion of negative weights.

4.3 Description of the Scenarios

Four different scenarios were considered for the possible configuration of the population of rotation groups for the rotation group sample design described in Section 2.2. The four scenarios provided different combinations of the rotation group size balance (good, poor) and of the correlation between the rotation group sizes M_{hi} and the survey variable y_{hi} (good, scattered). In the context of rotation group balance, "good" means that the rotation groups do not differ much in size, whereas "poor" means that they differ significantly. In the context of correlation, "good" means that the correlation between the survey variable and the rotation group size is quite high throughout the strata, whereas "scattered" means that it varies from low to high amongst the strata.

These scenarios represent possible configurations that will arise as the survey progresses through time. Scenario 1 reflects the survey at time of initial selection: for this case, the balance of rotation group sizes is good, and the correlation between rotation group sizes and the survey variable is good. Scenario 2 reflects the deterioration of the correlation (scattered) between the rotation group size and the survey variable as time progresses, due to dead units accumulating in the population. For this scenario, since the dead units have not been removed from the population, the balance in rotation group sizes is good, but the correlation between the survey variable and the rotation group size is weakened. Scenario 3 implies that removal of the dead population units may result in imbalance of the rotation group sizes (poor), but strengthening the correlation (good) between rotation group size and the survey variable. Finally scenario 4 represents the worst possible case, which is poor correlation between rotation group size and the survey variable, and poor balance in rotation group sizes.

Scenario 1 was constructed by varying the rotation group sizes and leaving the GBI values y_{hi} unchanged for all the rotation groups. The 16 rotation group sizes were varied by sorting them in ascending order of y_{hi}. Their size was set as follows. For rotation groups 1-4, 5-8, 9-12 and 13-16, the rotation group size was set to $0.22\, M_h/4$, $0.24\, M_h/4$, $0.26\, M_h/4$ and $0.28\, M_h/4$ respectively. The average correlation between the GBI and the rotation group sizes was 0.86, ranging from 0.69 to 0.96 at the individual stratum level. The average coefficient of variation of the rotation group sizes was 9.2%.

For scenario 2, the population units were randomly permuted and assigned systematically to one of 16 rotation groups, using the procedure described in Section 2.2.2. Approximately 20% of the population units were then randomly assigned a y-value of zero to represent a high proportion of dead units. The overall correlation between the GBI and the rotation group sizes was 0.11, ranging from -0.23 to 0.74 at the individual stratum level. The average coefficient of variation of the rotation group sizes was 4.1%.

For scenario 3 the procedure was similar to scenario 1 except that the rotation group sizes differed. For rotation groups 1-4, 5-8, 9-12 and 13-16, the rotation group size was set as 0.05 $M_h/4$, $0.20\, M_h/4$, $0.30\, M_h/4$ and $0.45\, M_h/4$ respectively. The overall correlation between the GBI and the rotation group sizes was 0.87, ranging from 0.70 to 0.96 at the stratum level. The average coefficient of variation of the rotation group sizes was 60.2%.

For scenario 4 a random rotation group size was assigned independently of the GBI values as follows. Suppose that for each stratum h, $a_h = \min\{M_{hi}\colon i = 1, \ldots, N_h\}$ and $b_h = \max\{M_{hi}\colon i = 1, \ldots, N_h\}$. For each stratum h, the size M_{hi}^* for rotation group i was set to r_h e_{hi} where e_{hi} is uniformly distributed on the interval (a_h, b_h). Here r_h is a scaling factor such that $M_h = \sum_{i=1}^{N_h} M_h^*$. The average correlation was 0, ranging from -0.49 to 0.56 at the stratum level. The average coefficient of variation of the rotation group sizes was 49.2%.

4.4 Discussion of Results

Based on the 4 scenarios described in the previous section, simulations were performed to compute the absolute relative bias (ARB), the absolute standard bias(ASB), the efficiency (EFF), and the proportion of weights less than or equal to 0. Those quantities were computed for each individual stratum and at the aggregate level. The results are given in Tables 1 to 3. Note that all of these results are presented as percentages.

In terms of absolute relative bias (ARB), as shown in Table 1, both the simple expansion and Mickey's estimator have no bias, as expected, neither at the overall nor at the stratum level. The separate ratio estimator displays the most absolute relative bias while the combined ratio estimator displays the least relative bias. For the biased estimators, the absolute relative bias increases as the coefficient of variation of the rotation group sizes increases, and the correlation between the rotation group sizes and the variable of interest decreases.

Turning to absolute standard bias (ASB), as shown in Table 2, the following observations can be made. The separate ratio estimator is unacceptable for most scenarios using this criterion. Its performance worsens as the variation in rotation group sizes increases, and as the correlation between the rotation group sizes and as the variable of interest decreases. The performance of the combined ratio estimator is acceptable, both at the aggregate and stratum level.

Table 1
Percentage Absolute Relative Bias ($\overline{\mathrm{ARB}}$)

Scenario	Aggregate Level		Stratum Level	
	Separate Ratio	Combined Ratio	Separate Ratio	Combined Ratio
1	1.27	0.07	1.31	0.11
2	0.02	0.01	0.24	0.05
3	2.88	0.14	3.19	0.29
4	5.51	0.22	5.72	0.30

Table 2
Percentage Absolute Standard Bias ($\overline{\mathrm{ASB}}$)

Scenario	Aggregate Level		Stratum Level	
	Separate Ratio	Combined Ratio	Separate Ratio	Separate Ratio
1	13.41	0.76	3.37	0.24
2	0.44	0.26	0.58	0.25
3	45.64	2.13	12.11	0.69
4	43.29	1.96	9.88	0.71

The behaviour of the estimators with respect to relative efficiency (EFF) is provided in Tables 3a and 3b. For Scenario 1, which represents good rotation group balance and good correlation, all the estimators are nearly equivalent, both at the aggregate and the stratum levels. For Scenario 2, which represents well balanced rotation groups and scattered correlation, the same conclusion holds. For Scenario 3, which represents poor rotation group balance and good correlation between the rotation group sizes and the survey variable, the ranking of the estimators at the aggregate level from highest EFF to lowest EFF is: i) the combined ratio, ii) the separate ratio estimator, iii) Mickey's estimator, and iv) the simple expansion estimator. For Scenario 4, which represents the worst in terms of rotation group balance and correlation between the rotation group sizes and the survey variable, the best estimator at both the aggregate and stratum levels is the simple expansion estimator. The combined ratio estimate is the next best choice.

Weights smaller than zero occured for the Mickey estimator in 2% of the cases.

In conclusion, given the above four scenarios, the combined ratio estimator is a reasonable choice for estimation for sub-annual surveys which use the rotation group design. The simple expansion estimator may also be considered on account of its simplicity. However, one should be aware of its poor conditional properties if the rotation group sizes are not balanced.

Table 3a

Percentage Relative Efficiency (EFF) at the Aggregate Level

Scenario	Simple Expansion	Separate Ratio	Combined Ratio	Mickey
1	100.0	108.0	107.9	107.3
2	100.0	100.2	99.8	100.1
3	100.0	148.3	160.3	143.5
4	100.0	74.3	92.3	84.3

Table 3b

Percentage Average Relative Efficiency ($\overline{\text{EFF}}$) at the Stratum Level

Scenario	Simple Expansion	Separate Ratio	Combined Ratio	Mickey
1	100.0	109.6	108.6	108.6
2	100.0	100.9	99.5	100.6
3	100.0	183.3	180.2	174.2
4	100.0	80.0	99.4	83.7

5. CONCLUSION

In this paper, we have presented a sample design which can accommodate the necessary requirements for a sub-annual business survey. These requirements have included initial sample selection, sample rotation and updating. Given this rotation group design, a number of estimation procedures have been considered and they have been evaluated via a simulation study. These estimation procedures are equivalent when the rotation group sizes are well balanced within each of the strata. In the case of unbalanced rotation group sizes, the use of the combined ratio estimator which used rotation group sizes as auxiliary information is recommended.

ACKNOWLEDGEMENTS

The authors would like to thank Lyne Guertin for the programming of the simulation study. We would also like to thank the referees and Professor J.N.K. Rao for valuable comments and constructive suggestions.

REFERENCES

BANKIER, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

BREWER, K.R.W., EARLY, L.J., and HANIF, M. (1984). Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference*, 10, 15-30.

BREWER, K.R.W., EARLY, L.J., and JOYCE S.F. (1974). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 232-239.

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

DALENIUS, T., and HODGES, J.L., Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

EARLY, L.J., and BREWER, K.R.W. (1971). Some estimators for arbitrary probability sampling. Master's thesis.

HAJÉK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley and Sons.

HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.

RAO, J.N.K., and KUZIK, R.A. (1974). Sampling errors in ratio estimation. *Sankhyā*. Series, C, 36, 43-58.

SUNTER, A.B. (1977). Response burden, sample rotation and classification renewal in economic surveys. *International Statistical Review*, 45, 209-222.