PROBLEMS ASSOCIATED WITH DESIGNING SUB-ANNUAL BUSINESS SURVEYS

M.A. Hidiroglou and K.P. Srinath, Statistics Canada
Business Survey Methods Division, 11th floor - R.H.Coats Bldg, Tunney's Pasture, Ottawa, Ontario, K1A 0T6
Statistics Canada

Key Words: Sampling Frame, Sample Rotation, Classification, Outliers

1. INTRODUCTION

Generally, business surveys are repeated monthly, quarterly or annually to produce estimates of totals, averages, ratios and estimates of change between two periods for several characteristics of interest. The problems associated with designing these surveys are common to most of the business surveys and tend to be more complex than the ones encountered in other surveys because of the highly dynamic nature of the sampling frame. Births and deaths of sampling units cause continuous additions and deletions to the frame and classification information on the frame (industry, size etc.) changes continuously with time. Sample selection and maintenance procedures have to take into account all these changes. The need to produce reliable estimates of change between periods and also reduce response burden requires some form of sample rotation. Estimation procedures should take into account outliers which could distort the estimates of totals. A brief discussion of some of the problems can be found in Finkner and Nisselson (1978), Sanyal and Sinha (1977), Konschnik et al. (1985) and Srinath (1987). It is important that satisfactory solutions to these problems are found as otherwise, the estimates could be subject to large biases or unacceptably high variances.

In this article, a discussion of the some of the more important problems is given and practical solutions to these problems are proposed. The problems relating to construction and maintenance of sampling frames for surveys of businesses are detailed in Section 2. Methods of stratification, allocation and sample size determination for populations having highly skewed distributions are outlined in Section 3. Section 4 gives some methods of sample rotation and procedures for selection of births and elimination of deaths which attempt to keep the sample representative of the changing population. The difficult problem of implementing classification changes to the units both in and out of the sample without affecting the estimates of change is discussed in Section 5. Sections 6 and 7 deal briefly with the problems of edit and imputation and estimation as well as dealing with outliers. Finally, some concluding remarks are given in Section 8.

2. FRAME REQUIREMENTS

List frames are used extensively in business surveys and it is not uncommon to employ both area and list frames. It is not easy, however, to construct a list of sampling units for purposes of drawing a sample. Generally, there is no single source for a complete and current list of businesses with correct names, addresses and standard industrial classification information. Therefore, the sampling frames for business surveys have the usual inaccuracies such as duplication, incompleteness, errors in classification of units and presence of extraneous units.

Businesses are composed of legal entities and operating entities where the legal entity is the legal representation of the business while the operating entity organizes and controls the production of goods and services. Although the majority of businesses have a simple structure consisting of one legal entity owning and controlling one operating entity, larger businesses can have complex structures and relationships between legal and operating entities, the complexity increasing with the size of the operating entity. A detailed description of possible complexities in structures can be found in Cuthill (1989).

For purposes of sampling and data collection using a frame of businesses with complex structures, it is necessary to derive a frame of statistical units based on the relationship between operating and legal entities. But it is by no means easy or inexpensive to delineate statistical units, let alone maintain the delineation over time. The delineation is dependent on the knowledge of the entire structure of the business which is obtained through an expensive operation known as "profiling". Profiling has to be done on a reasonably continuous basis as businesses change in composition through mergers and acquisitions. Because of resource constraints it is usually impossible to profile the entire population of businesses and to keep the profiles updated. Therefore there will always be inaccuracies on the frame.

Another difficulty is the definition of a sampling unit. It is not always clear what this should be because of different data requirements to be met by a single survey of both large and small businesses. For example, if a survey is to collect data on employment, earning and hours, as well as on total employee compensation, the sampling unit could be an establishment, for the former purpose and the company for the latter purpose. The data on employment is available at the establishment level and the data on employee compensation is available only at the company level. Therefore, a survey could involve more than one type of unit. For further examples, and a detailed explanation of the problems and solutions see Colledge (1987).

Maintaining a frame of businesses over time is complicated because of the dynamic nature of the frame. Mergers, acquisitions, changes in ownership and reorganizations, etc., necessitate setting up rules for handling changes. These rules for maintenance have to be set up in a way as to minimize the bias in the estimates.

3. STRATIFICATION, ALLOCATION AND SAMPLE SIZE DETERMINATION

The stratification of a business universe is normally 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). A primary stratum is

the cross-classification of industry and geography regions for which estimates are required. Within these primary strata, further strata are formed using the size measure of the units. Efficient sampling of highly skewed populations such as those of businesses require that the units within each primary stratum be stratified into a take-all stratum and a number of take-some strata. Units belonging to the take-all stratum are selected with certainty, whereas units in the take-some strata are selected using a probability mechanism.

The stratification of a highly skewed population into two strata usually requires further stratification of the take-some stratum, and efficient allocation requires that Neyman allocation be used. This can be achieved using the Dalenius-Hodge (1959) cum$\sqrt{f}$ rule or Hansen's (1953) cum$\sqrt{x}$ rule. Further gains can be achieved if one simultaneously stratifies the population into a take-all stratum and a number of take-some strata. Lavallée and Hidiroglou (1988) provided such a procedure for minimizing the overall sample size given a fixed coefficient of variation and Neyman allocation of the sample to the take-some strata. Their algorithm is a modification of Sethi's (1963) method for stratifying a population.

The following is an extension to this algorithm for any allocation scheme. Consider a finite ordered population of n units, $y_{(1)}, y_{(2)},...,y_{(N)}$, with $y_{(i)} \le y_{(i+1)}$ for i = 1, 2,..., N-1. This population is to be stratified into L strata. The number of units in each stratum is denoted as N_h, h = 1, 2,..., L. The sampling scheme calls for n_h units to be drawn without replacement using simple random sampling from the N_h population units in stratum h where the L-th stratum is the take-all stratum and the others are the take-some strata.

An estimator of the total $Y = \sum_{h=1}^{L} \sum_{j=M_{h-1}+1}^{M_h} y_{(j)}$ is given by

$$\hat{Y} = \sum_{h=1}^{L-1} \frac{N_h}{n_h} \sum_{j=m_{h-1}+1}^{m_h} z_j + \sum_{j=M_{L-1}+1}^{M} y_{(j)} \qquad (3.1)$$

where $M_h = \sum_{i=1}^{h} N_i$ $(h=1, 2, \ldots, L)$, $M_o = 0$, $m_h = \sum_{i=1}^{h} n_i$, $m_o = 0$,

and $y_{M_{h-1}+1} \le z_j \le y_{M_h}$ for j = $M_{h-1}+1,...,M_h$ (h = 1, 2, ..., L-1).

For a given allocation scheme, let a_h $(\sum_{h=1}^{L-1} a_h = 1)$ be the proportion of units to be allocated to the h-th stratum (h = 1, 2, ..., L-1). Assuming a desired level of precision c (coefficient of variation) the overall sample size is

$$n = N_L + \frac{\sum_{h=1}^{L-1} N_h^2 S_h^2 / a_h}{c^2 \hat{Y}^2 + \sum_{h=1}^{L-1} N_h S_h^2} \qquad (3.2)$$

where S_h^2 (h = 1, 2,..., L-1) is the population variance of stratum h. The resulting sample sizes for each take-some strata will be $n_h = (n-n_L)\, a_h$. The objective is to determine boundaries $b_1, b_2,..., b_{L-1}$ (where $y_{(1)} < b_1 <...< b_{L-1} < y_{(N)}$) such that the overall sample size n is minimized.

Assume that the distribution of the population units can be represented by a continuous density function f(y). Given this continuous representation equation (3.2) can be rewritten as

$$n = N W_L + \frac{N\left(\sum_{h=1}^{L-1} W_h^2 \sigma_h^2 / \gamma_h\right)\left(\sum_{h=1}^{L-1} \gamma_h\right)}{N c^2 \mu^2 + \sum_{h=1}^{L-1} W_h \sigma_h^2} \qquad (3.3)$$

where

$$\gamma_h = W_h^{2p_1} \mu_h^{2p_2} \sigma_h^{2p_3}, \quad (p_i \ge 0;\ i=1, 2, 3),$$
$$W_h = \int_{b_{h-1}}^{b_h} f(y)\, dy, \qquad (3.4)$$

$$\mu_h = \int_{b_{h-1}}^{b_h} y f(y)\, dy / W_h, \qquad (3.5)$$

$$\sigma_h^2 = \int_{b_{h-1}}^{b_h} y^2 f(y)\, dy / W_h - \mu_h^2 \qquad (3.6)$$

$$\mu = \int_{b_o}^{b_L} y f(y)\, dy \qquad (3.7)$$

For h = 1, 2, ... ,L with $b_0 = -\infty$ and $b_l = \infty$. Different combinations of the p_i's assigned to γ_h yield power allocations (when $p_1 > 0$, $p_2 > 0$, and $p_3 = 0$), as discussed in Bankier (1988), the usual Neyman allocation (when $p_1 = p_3 = 1$ and $p_2 = 0$), as well as combinations of those allocations. The optimum boundaries are obtained by differentiating equation (3.3) with respect to b_h, h = 1, 2, ... ,L-1. This differentiation yields quadratic equations in the b_h's of the form

$$[F(G_{1h}-G_{1,h+1})]\, b_h^2 + [F(G_{2h}-G_{2,h+1}) + 2AB(\mu_h-\mu_{h+1})]\, b_h + [F(G_{3h}-G_{3,h+1}) - AB(\mu_h^2 - \mu_{h-1}^2)] = 0 \qquad (3.8)$$

for h = 1, 2,...,L-2

and

$$[FG_{1,L-1}]\, b_{L-1}^2 + [FG_{2,L-1} + 2AB\mu_{L-1}]\, b_h + [FG_{3,L-1} - AB\mu_{L-1}^2 - F^2] = 0 \qquad (3.9)$$

for h = L-1,

where

$$G_{1h} = B p_3 Z_{3h} + 2A(1-p_3) Z_{4h}$$
$$G_{2h} = 2[B p_2 Z_{2h} - G_{1h} \mu_h]$$
$$G_{3h} = 2q(B Z_{1h} - A Z_{5h}) + G_{1h}(\mu_h^2 + \sigma_h^2)$$
$$Z_{1h} = W_h^{2q-1} (W_h\mu_h)^{2p_2} (W_h\sigma_h)^{2p_3}$$
$$Z_{2h} = W_h^{2q} (W_h\mu_h)^{2p_2-1} (W_h\sigma_h)^{2p_3}$$
$$Z_{3h} = W_h^{2q} (W_h\mu_h)^{2p_2} (W_h\sigma_h)^{2p_3-1} \sigma_h^{-1}$$
$$Z_{4h} = (W_h\sigma_h)^{1-2p_3} W_h^{-2q} (W_h\mu_h)^{-2p_2} \sigma_h^{-1}$$
$$Z_{5h} = (W_h\sigma_h)^{2-2p_3} W_h^{-2q-1} (W_h\mu_h)^{-2p_2}$$
$$Z_{6h} = (W_h\sigma_h)^{2-2p_3} W_h^{-2q} (W_h\mu_h)^{-2p_2-1}$$
$$A = \sum_{h=1}^{L-1} W_h^{2p_1} \mu_h^{2p_2} \sigma_h^{2p_3}$$
$$B = \sum_{h=1}^{L-1} (W_h \sigma_h)^2 W_h^{-2p_1} \mu_h^{-2p_2} \sigma_h^{-2p_3}$$

and

$$q = p_1 - p_2 - p_3.$$

Labelling the coefficients of b_h^2 as α_h, the coefficient of b_h as β_h and the remaining terms as δ_h, equations (3.8) and (3.9) can be represented as quadratic equations of the form $\alpha_h b_h^2+\beta_h b_h+\delta_h=0$. Since α_h, β_h and δ_h involve some knowledge of f(y), the approximate density function is required to solve (3.8) and (3.9). This can be achieved by replacing the quantities in (3.5), (3.6) and (3.7) by their finite population counter parts. These are respectively:

$$W_h = N_h/N \tag{3.10}$$

$$\bar{Y}_h = \sum_{j=b_{h-1}+1}^{b_h} y_{(j)} / N_h \tag{3.11}$$

and

$$S_h^2 = \sum_{j=b_{h-1}+1}^{b_h} (y_{(j)}-\bar{Y}_h)^2 / (N_h-1) \tag{3.12}$$

for h = 1, 2,...,L.

Substituting these quantities into α_h, β_h and δ_h, the following algorithm, as suggested by Sethi (1963), can be used to find the optimal boundary points b_h, h = 1, 2,...,L-1.

STEP 0: Sort the population $y_1,...,y_N$ in ascending order and set $b_0 = y_{(1)}$ and $b_L = y_{(N)}$.

STEP 1: Start with some arbitrary boundaries such that $b_0<b_1'<...<b_{L-1}'<b_L$.

STEP 2: Calculate the proportions W_h', the mean $\bar{Y}'_h$ and the variance S'^2_h (from equations (3.10), (3.11) and (3.12) respectively) based on the boundaries in step 1.

STEP 3: Replace the initial set of boundaries by $b_1'',...,b_{L-1}''$.

where $b''_h = \dfrac{-\beta'_h+\sqrt{\beta'^2_h-4\ \alpha'_h\ \delta'_h}}{2\ \alpha'_h}$, h = 1, 2, ..., L-1

STEP 4: Repeat steps 2 and 3 until two consecutive sets are either identical of differ by negligible quantities, i.e. $\max_{h=1}^{L-1} | b''_h-b'_h | < \epsilon$ for some $\epsilon>0$.

It can be proved that the sign before the square root (√) is positive because b_h' lies between $\bar{Y}'_h$ and $\bar{Y}'_{h+1}$.

The allocation can be adjusted to achieve the desired minimum sample and/or maximum weights for each secondary stratum. It should be noted that the stratification and allocation are done to achieve the desired efficiency for estimating totals and averages. The allocations generally lead to the smallest sampling fraction in the stratum of smallest businesses.

4. SAMPLE ROTATION, BIRTHS AND DEATHS

In this section, methods for initial sample selection, rotation, selection of births and elimination of deaths are given. These methods take into account the dynamic nature of the business universe and also ensure that the sample on any occasion reflects the structure of the population on that occasion and so continues to be representative of the population. Sample rotation or partial replacement of the sample at each occasion is done in business surveys primarily to reduce response burden. Partial replacement of the sample as opposed to keeping the same sample or taking a completely new sample on each occasion, also helps to get better estimates of both change and annual averages. Generally, sample rotation has to be done under certain constraints. For example, in a monthly survey, it may be important to keep businesses that have been rotated out of the sample from rotating back into the sample for at least 12 months thereafter. That is, they are not eligible for selection at least for 12 months after being rotated out. This 'time-out' requirement might dictate the length of time a sampling unit is retained in the sample or the 'time-in'. For example, if it is also desired to keep businesses in the sample for 12 months, then in some strata units may have to be kept in the sample for more than 12 months in order to satisfy the time-out constraint. This is considered desirable since it is more of a response burden for a business to come back into the survey within a certain number of months than to continue reporting each month for a number of months. Three possible methods of rotation that attempt to meet these constraints are described below.

4.1 Panel Sampling

The selection procedure requires that the sampling units in the population in each stratum be grouped into a certain number of clusters or panels and then a sample of panels be selected. The number of panels in and out of the sample depend on the sampling fraction in the stratum, and also time-in and time-out constraints. If there is no time-out constraint, then the number of panels is simply determined by multiplying the inverse of the sampling fraction and the number of occasions that the unit is to be in the sample. However this straightforward procedure cannot ensure that the units stay out of the sample at least for a certain period after they rotate out. The actual procedure for ensuring time-in and time-out constraints is as follows.

Let the number of panels in the population in the stratum be P and the number of panels in sample be p. Let N denote the population size, n the sample size, TI the desired number of occasions a unit is to be in the sample, TO the minimum number of occasions a unit is required to stay out of the sample and f the desired sampling fraction in the stratum. The following method of determining P and p will ensure that the units stay out of the sample at least for TO occasions. Compute

$$X = int\left|TI\frac{1-f}{f} + 0.5\right|$$

If X > TO, then the number of panels in sample is p = TI and the number of panels out of sample is P-p = x.
If X < TO, then the number of panels in sample is

$$p = int\left|TO\frac{f}{1-f} + 0.5\right|$$

and the number of panels out of sample is P-p = TO. Note that P and p are determined in such a way that p/P = n/N at the time of initial selection.

The method of allocation of units in the population to the panels depends on whether the population size is greater

or less than the predetermined number of panels. The panels are first numbered 1, 2, 3, ... P. This ordering is called "rotation ordering". This is the order in which the panels will be selected in the sample and also rotated in and out of the sample. A random permutation of this ordering is called "assign ordering". If N > P, say N = mP + r where m > 1 and r ≥ 0. The first unit is assigned to the first panel and the second unit to the second panel according to the "assign ordering", and so on, the Pth unit going to the Pth panel. The (P + 1) th unit is again assigned to the first panel so that the first r panels have m + 1 units each and the next P - r panels have m units each. When N < P the panels which will be non-empty are first determined thus: a random number is selected between 1 and P/N = k (say). Let this number be r. Then the r, r + k, r + 2k ..... r + (N - 1)k-th panels in the rotation ordering are selected to be non-empty. For the purpose of assigning the units, a random permutation of the N units is used. The first random unit in the population is assigned to the first non-empty panel in the assign ordering, the 2nd unit to the 2nd panel and so on and finally the Nth unit assigned to the Nth panel in the assign ordering.

The panels are selected in the sample using rotation ordering. That is for the first occasion panels numbered 1, 2, 3, ... p in the rotation ordering are selected in the sample. For the case N > P all the selected panels will be non-empty where as for the case N < P panels will have either 0 or 1 unit each.

4.1.1 Sample rotation:

Sample rotation using panels is simple. As noted earlier, the panels numbered 1, 2, 3, ... p in the rotation ordering are included in the sample on the first occasion. For the second occasion, panel 1 is dropped and the panel numbered p + 1 is included.

4.1.2 Selection of Births:

Births occur as a result of a new business activity being started or because of restructuring of an existing business such as a change of ownership or a change of industrial activity that brings the business from out of scope into in scope for the survey. Births are subject to stratification like other survey units. All births falling in a take-all stratum are included in the sample with certainty. For allocating the take-some births to the panels, assign ordering is used each month. On the second occasion, births are assigned starting from the (r + 1)th panel in the assign ordering for the case N > P. On each occasion the panel to which the last birth was assigned is noted and subsequent births are assigned to panels starting from the panel next to it. For the case N < P births are assigned only to the N panels again using the assign ordering such that each panel gets one birth in a sequential manner. Note that the empty panels are never assigned births and therefore remain empty for all occasions.

4.1.3 Advantages and Disadvantages

Panel sampling has the advantage of being operationally very simple. The procedure is unbiased being a simple random sample of panels and the sampling of births presents no difficulties. However, there is a possibility that the panel sizes could become unequal over time, because of deaths, although initially all panels differ in size by one unit at most. Therefore, it may not be possible to guarantee either minimum sample size or minimum sampling fraction in terms of the number of sampled units. Since the sample size on each occasion could vary, it may not be possible to meet the reliability requirements for the estimates. Also, the scheme does not easily permit the control of overlap between surveys.

4.2 Collocated Sampling:

The procedure of selection under this method is very simple and consists of the following steps.

(i) Assign an equispaced sample selection number SSN(i) to the ith unit in the population as follows

$$SSN(i) = (R + i - 1)/N \qquad i=1, 2, 3, \ldots, N$$

where R is a uniform random number between 0 and 1 generated for each stratum and where N is the population size of the stratum.

(ii) All units whose sample selection numbers lie within the interval (0,f) where f is the desired sampling fraction in that stratum are included in the sample.

4.2.1 Sample Rotation

Rotation of the sampled units is achieved by simply shifting the sampling interval (0,f). The amount of shift, say "s", depends on the "time-in-sample" and "time-out of sample" constraints. After the shift, all the units which have sample selection numbers with the new interval (0+s, f+s) are included in the sample. That is, the units in the interval 0 to s are rotated out and units in the interval f to f+s are rotated in. The shift in the interval is calculated as

$$s = min\,(f/time\text{-}in,\ 1-f/time\text{-}out).$$

4.2.2 Selection of Births

On each occasion births are assigned an equispaced sample selection number independent of the numbers assigned to older units. That is, births are equispaced among themselves. All the births which have sample selection numbers within the sampling interval for that occasion are included in the sample. Equispacing of births ensures that there is no overselection or underselection of births. The expected number of births in the sample is equal to f times the number of births in the population.

A similar procedure is followed for subsequent births. Equispacing births each month among themselves avoids clustering of births in the sample and ensures a proper representation of births each month. Moreover, a common weight could be used at the estimation stage since all units are sampled at the same rate.

4.2.3 Advantages and Disadvantages:

As noted earlier, the procedure is operationally very

simple. Another advantage of this method is that it ensures the desired sample size. It also facilitates the control of response burden over several surveys if this is desired later. This can be done by reserving specific sampling intervals for each survey. The method allows for changing sampling fractions. For a description of this see Tambay (1988). Though selections are made independently for births each month, the sample is treated as a simple random sample for estimation purposes. This should not cause any bias as births are sampled at the same rate. A disadvantage of the method is there is no equispacing of births which occur in different months, though they are equispaced each month.

4.3 Rotation Group Method:

In this method only the sampled units in a stratum are divided into what are known as rotation groups. The number of groups depends on the time-in sample period. For example, in a monthly survey, if a sampling unit is desired to be in the sample for 12 months then there will be 12 rotation groups in each stratum. The rotation group labels 1 to 12 indicate the month in which the units (other than births) rotated into the sample. Rotation group 1 contains units which entered the sample in January, rotation group 2 consists of units which entered the sample in February and so on, rotation group 12 consisting of units which rotated into the sample in December. At the time of initial selection (on the first occasion) the sampled units are randomly assigned to the rotation groups in such a way that the size of the rotation groups differ from each other by one unit at most. The non-sampled units fall into two lots known as Lot I and Lot II. Lot I consists of units which have not yet been selected in the sample as well as units which have completed their time-out of sample period and therefore are eligible for selection again. Lot II consists of units which have rotated out of the sample but have not yet completed their time-out of sample period. Units in Lot II are again assigned to a certain number of groups, the number depending on the time-out of sample period. For example, if it is desired to keep units out for 12 months then there Lot II will consist of 12 groups labelled 13 to 24. Units rotating out of the survey in January go into group 13 in Lot II and are not eligible for selection till January of next year.

4.3.1 Sample Rotation

At the time of monthly sample selection and rotation, say in month 1 of the next year, all units in group 13 in Lot II are transferred first to Lot I. The units in rotation group 1 in the sample are transferred to group 13 in Lot II. After this a sample of units is selected from Lot I and is placed in rotation group 1. The number of units selected to be placed in rotation group 1 is determined so that the sampling fraction for that stratum determined at the time of initial selection is kept constant in order to obtain a certain reliability for the estimate. The process of exchanging units between groups is done in order to satisfy the time-in and time-out constraints and also to ensure that the sample is representative of the population. A more detailed description of the method can be found in Schiopu-Kratina and Srinath (1991).

4.3.2 Selection of Births

Births are selected in the sample using the same sampling fraction determined at the time of initial selection. That is, if B births are added to the stratum for any occasion then $b = fB$ births are selected where f is the sampling fraction in the stratum.

The selected births are assigned to the rotation groups in the sample at random in such a way that each rotation group differs in the number of births it gets, at most by one. The non-selected births are assigned to Lots I and II in proportion to the number in each group. The births falling in Lot II are again assigned to the groups at random. The assignment of non-selected births as described above is to ensure that the sample continues to reflect the structure of the population in terms of births and other units.

4.3.3 Advantages and Disadvantages:

A disadvantage of this procedure is that it is more complex than the two described above. The number of units to be selected at each occasion will have to be computed. The selection of births and the allocation of non-selected births is more involved. The procedure is more flexible than the panel approach in the sense that it is easier to change sampling fractions mid-stream. It also ensures minimum sampling fraction and sample size requirements without violating the time-in and time-out constraints.

4.4 Elimination of Deaths:

It is well known that the lag between the time a business ceases to operate and its removal from the business frame is considerable. It is even longer than the birth-time lag mentioned earlier. Therefore, businesses are always sampled from a frame that contains a large number of "out of business" or extraneous units. The burden of retaining an increasing number of inactive units on the frame could lead to apparent nonresponse to the survey, at least for the initial occasions where the reason for no response is not yet established. Also, the estimates based on samples drawn from such a population are likely to have a large variance due to the fact that the population contains a high proportion of zero observations. This invariably leads to estimates of level and change not meeting the reliability requirements. Ideally, all such units should be eliminated from the sampling frame before the sample is drawn. This is difficult to achieve. It is also difficult to guess the number of such businesses on the frame especially when the economy is on a downward trend.

Identification of businesses which are no longer operating takes place primarily through the sample. Even here, there is a time-lag and the business may be treated as a nonrespondent in the initial period of its selection leading to overestimation of totals with traditional imputation methods. These units also add to the cost of the survey as they require rigorous follow-up. Eliminating the inactive units from the sample without a corresponding elimination of such units in the population will lead to a bias in the estimates, if the weights involve the known population and sample sizes.

The presence of dead/inactive units at the time of sampling necessitates the determination of procedures for

their removal from the frame. One simple unbiased procedure in a continuing survey is to remove "deaths" from the sample only if they are identified as such by a source which is independent of the sampling process. That is, this source identifies units as "dead" irrespective of whether the units are in the sample or not. But if such updating occurs infrequently, deaths tend to remain on the frame for a long time, and this is a clear disadvantage because of the problems mentioned above. This infrequent removal of deaths may also cause blips in the estimates due to a disproportionate number of deaths in the sample relative to the nonsampled portion of the population. This is especially true if the weight used for estimation is a function of the number of units in the sample and the number of units in the population. Some kind of balancing in the number of deaths in and out of the sample may be required to keep the weights constant in order to avoid this artificial change in the estimates of trends. The identification of the source as strictly independent of the survey may be difficult if the frame is being updated through several sources and is being used by several surveys.

5. CHANGES IN CLASSIFICATION INFORMATION

As noted earlier one important feature of repeated business surveys is the dynamic nature of the sampling frame. Businesses are constantly coming into existence, going out of business, merging and splitting. In addition to this, there are changes to the classification information. These changes include changes in industry, size and geography and can occur in either the sampled or the non-sampled portion of the population. Changes in the industry classification, location or size could be real reflecting the change in the activity of the business, its location or size. A change in classification could also come about because the business was originally misclassified. These changes, are detected for in-sample units more often than for out of sample units. In a one time survey, if a change in the classification information is noticed after the selection of the sample, this is handled through domain estimation. That is, the weight originally assigned to the sampled unit is retained for estimation purposes but the unit is assigned to the new stratum for tabulation. In repeated surveys, units have to be reclassified at some point in time, though domain estimation can be used for several occasions.

Classification changes can be implemented, if there is a source independent of the sampling process which identifies changes in the information relating to all the units in the population. This is unrealistic to expect in the case of large populations. If such a source is not available, then these changes are stored and implemented at the end of a given time period. This might cause an artificial blip in the estimates in addition to necessitating a selection or a partial selection of a new sample. The selection of a new sample may distort the rotation scheme. Generally, units which need to change strata are treated as births and deaths. That is, units which now do not belong to the stratum in which they were originally selected are treated as deaths in that stratum and removed and allocated as births in the new stratum to which they now correctly belong. In the absence of an outside source which updates the sampling frame on a universal basis, rules may have to be devised in order to implement classification changes such that biases in the estimates of level and change are kept to a minimum.

6. EDIT AND IMPUTATION

The problems which will be addressed in this section are the editing and imputation of data for the sampled units. The nature and quality of responses can affect the data consistency (quality) over a given time period. The reporting unit may report the data faithfully with no dramatic departure in continuity ("smoothness") as time progresses, or, there may be questionable jumps between two time periods. The reporting unit may not report all the requested data items: this is known as partial non-response. The reporting unit may report data sporadically with breaks causing total non-response for some periods. These cases can occur simultaneously in a periodic survey.

6.1 Statistical Edits:

Statistical edits are used to isolate reporting units which may report some of their quantitative data fields in an inconsistent manner either from time period to time period or within a specific time period. Units with unusually high or low values, as compared to the data reported in the previous time period, will be termed "outliers" if they are markedly different from other units. The identification of "outliers" is extremely important in an ongoing survey for two reasons. First, they influence estimates from the data set, such as those for totals. Second, since the imputation of quantitative data for non-responding units is usually based on trends, means or medians, the removal of outlier units from the computation of these trends, means or medians, will produce statistics that are not contaminated with these observations.

Units which have data that are not consistent within a given time period are found using consistency edits. For a given unit i and time period t, let $\mathbf{x}_i(t)$ represent the vector of data to be collected. The vector $\mathbf{x}_i(t)$ may be decomposed into elementary vectors for which independent editing and imputation are required. That is,

$$\mathbf{x}_i(t) = \left(\mathbf{x}_i^{(1)}(t), \ldots, \mathbf{x}_i^{(p)}(t)\right) \qquad (6.1)$$

where $\mathbf{x}_i^{p}(t) = \left(x_{i1}^{(p)}(t), \ldots, x_{ik_p}^{(p)}(t)\right)$

for i = 1, ...,n ; p = 1, ...,P ; t = 1, ...,T and k_p is the number of variables in the p-th elementary vector.

For each elementary vector $\mathbf{x}_i^{(p)}(t)$, the consistency edits may be represented as

$$\mathbf{A}^{(p)}\left(\mathbf{x}_i^{(p)}(t)\right)' < C^{(p)}$$

where $A^{(p)}$ is a ℓ_p by k_p matrix representing the rules that the elements of the elementary vector $x_i^{(p)}(t)$ must obey and $C^{(p)}$ is a 1 by ℓ_p vector which represents the constraints. This formulation allows one to define consistency edits for both qualitative and quantitative variables.

The isolation of units which have data which differ markedly in behaviour between two time periods should be done bearing in mind the following. Firstly, their behaviour must be different from other "similar" units. Secondly, importance should be given to those units with the greatest impact on changes: these will usually be the largest units with respect to size. Lastly, the method for isolating these units should be relatively simple to implement. A procedure which has been found to be satisfactory for satisfying these

conditions in practice was given by Hidiroglou and Berthelot (1986). This procedure is developed as follows. Define a combination of the change and magnitude of data as

$$e_i = s_i\{\max(x_i(t), x_i(t+1))\}^u \qquad (6.2)$$

where

$$s_i = \begin{cases} 1-r_M/r_i, & \text{if } 0<r_i<r_M \\ r_i/r_M-1, & \text{if } r_i>r_M \end{cases}$$

$$r_i = x_i(t+1)/x_i(t)$$

$$r_M = \text{median}\{r_i : i=1,\ldots,n\}$$

and $0 \leq u \leq 1$. The e_i's are referred to as effects and u as the importance associated with the magnitude of the data. The parameter ucontrols the shape of the curve defining the upper and lower boundaries. The effect of increasing uis to attach more importance with fluctuations associated with the larger observations. Outliers are those units whose associated effect e_i lies outside the interval $(e_m - cd_{q1}, e_m + cd_{q3})$ where $d_{q1} = \max(e_m - e_{q1}, |ae_m|)$ and $d_{q3} = \max(e_{q3} - e_m, |ae_m|)$. Here e_{q1}, e_m and e_{q3} are respectively the first quartile, the median and the third quartile of the e_i's. The purpose of the ae_m term is to avoid difficulties which arise when $e_m - e_{q1}$ or $e_{q3} - e_m$ are too small. A value of 0.05 for a has proved to be adequate in practice. The parameter c controls the width of the acceptance interval.

6.2 Imputation

In most business surveys, subject to budget constraints, non-responding units are followed up in order to improve the response rates. This follow-up is usually carried out by mail in the case of the smaller to medium size non-responding units and by telephone for the larger units. Although this follow-up improves response rates, there will be nevertheless a group of non-responding units which may be classified into either hard-core or late respondents. Hard-core non-respondents are units which require a great deal of follow-up in order to respond, if at all. Late respondents are units which respond late with respect to the survey's reference period, either because they do not mail back their questionnaire on time or because they need to be prompted by follow-up questionnaires. These non-responding units must therefore be imputed in order to make up for their contribution to the estimates that need to be produced within established dead-lines/dates limits. It must be noted that these imputation procedures can also be used to generate values for units declared as outliers, if no valid explanation can be provided for their unusual values in the estimates. The resulting imputed values can be used in lieu of these outlying observations.

Units with no response whatsoever, will be termed as total non-respondents and those with partial non-response will be termed as partial non-respondents. The following desirable features relate to an imputation system which must impute for total non-response on a monthly basis. The imputation cell, the level at which the computation of trends, means or medians is performed, will usually correspond to the level of stratification of the sample. A minimum number of units must participate in the computation of these trends means or medians. Otherwise, the imputation cells will be automatically collapsed (using a pre-determined pattern, until the minimum requirement has been satisfied. The collapsing pattern is worked out by grouping cells which have similar behaviour. Status codes are used to control and keep track of the imputation process. For example, values are not imputed for seasonal units during the period that they are not operating, for units that are temporarily out of business, and for inactive units. The type of imputation used (trend or mean) is recorded. The most reasonable imputation procedure (trend or mean) under the existing data configurations is automatically determined. For non-responding units which are new (births or units rotating into the sample) to the survey, the data will be imputed using the mean or median of responding new units in the cell. For non-responding units which have participated in the survey for more than one occasion, one of the following imputation procedures is used depending on the availability of data: i) trends (month over month, quarter over quarter, or year over year for the same month) with the most recent trends being given the priority, and (ii) imputing means or medians.

Monthly trends are applied to units which have data (response or imputed) in the month prior to the one to be imputed. Annual trends are used mostly for units which are seasonal and which fail to provide a response as they emerge from their out-of-season period. Imputations based on the trends are obtained by multiplying the trend by the unit's last month or last year value. In the event that trends cannot be applied, the mean or median of the cell is used.

7. ESTIMATION

Generally, there is a requirement to produce unbiased (or nearly unbiased) estimates along with the associated measures of reliability (coefficients of variation). The building blocks for aggregation are the strata from which samples have been selected. It is at this level that basic sampling weights are computed. Domain estimation is used to produce estimates. A domain can span across all the sampling strata or it can be a subset of a stratum. A desirable feature of the estimates is that the sum of any domain set must always add up to the domain defined as their union. This holds provided that the elements of the domain set are mutually exclusive. Domain estimation automatically takes into account units which have changed their classification (industry or size) since the time of sampling.

Sampling weights that do not incorporate available auxiliary information lead to estimates that are unconditionally unbiased but which can be conditionally biased or sometimes inefficient. Nearly conditionally unbiased estimation can be obtained by using ratio estimation.

Variance estimation must reflect the sample design as well as the imputation and estimation methods used. In the case of imputed data, serious underestimation of the variance will occur if they are treated as response data in the usual variance procedures. Särndal (1990) and Rao (1991) have proposed procedures to take imputation into account for variance estimation. Their methods only deal with one type of imputation at a time and need to be extended to account for data which may have been imputed using a mixture of imputation procedures.

The detection and treatment of units which dominate

the estimates for a given level of aggregation remains an open problem. The main questions with respect to this problem are as follows: i) at what level of aggregation should their detection occur, ii) how robust are the estimates to the assumption that they represent unique observations in the population?, iii) how much bias is acceptable? and iv) how much discontinuity are we willing to accept to the published results between survey occasions? The impact of such units can be reduced by i) either reducing their weight to one and subsequently modifying the weights of the remaining units in the stratum by ensuring that the sum of the weights over all units add up to the stratum population size (Hidiroglou-Srinath 1981) or by ii) Winsorizing the observations as in Fuller (1991). The Winsorization effectively brings back the values of influential units to a boundary which is determined from the estimated sample distribution. Both these methods lead to negative bias in the estimates.

8. CONCLUDING REMARKS

From a discussion of the problems described in section 2, it is easy to conclude that maintaining a current and correct list of businesses for purposes of sampling is a difficult and an expensive operation. It is difficult to eliminate the problems of undercoverage and duplication. Attempts must be made to at least measure these deficiencies in the frame in order to assess their impact on the estimates. Maintenance of the sample after classification changes require solutions which are operationally simple and at the same time minimize the bias in the estimates. The problem of determining and then dealing with outliers in surveys has no completely satisfactory solution. Solutions may have to be specific to surveys. In general, there is a need to provide solutions which strike a balance between being operationally feasible and simple and getting unbiased and efficient estimates.

REFERENCES

Bankier, M.D. (1988), "Power Allocations: Determining Sample Sizes for Sub-National Areas," the American Statistician, 42, 174-177.

Colledge, M.J. (1987), "Use of Administrative Data in the Business Survey Redesign Project," in Proceedings of the International Symposium on Statistical Use of Administrative Data, Statistics Canada, pp. 153-167.

Cuthill, I.M. (1989) "The Statistics Canada Business Register," in Proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census, pp. 69-86.

Dalenius, T. and Hodges, J.L. Jr. (1959), "Minimum Variance Stratification," Scandinavisk Aktuarietidskrift, 54, 88-101.

Finkner, A.L. and Nisselson, H. (1978), "Some Statistical Problems Associated with Continuing Cross-sectional Surveys," in Survey Sampling and Measurements, N.K. Namboodiri (ed), Academic Press, New York.

Fuller, W.A. (1991), Simple Estimators for the Mean of Skewed Populations," Statistica Sinica, 1, 137-158.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), "Sample Survey Methods and Theory, Vol. 1, New York: John Wiley.

Hidiroglou, M.A. and Berthelot J-M. (1986), "Statistical Editing and Imputation of Periodic Business Surveys," Survey Methodology, 12, 73-83.

Hidiroglou, M.A. and Srinath, K.P. (1981), "Some Estimators of Population Total from Simple Random Samples containing Large Units," Journal of the American Statistical Association, 76, 690-695.

Konschnik, C.A., Monsour, N.J. and Detlefsen, R.E. (1985), "Constructing and Maintaining Frames and Samples for Business Surveys," American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 113-122.

Lavallée, P. and Hidiroglou, M.A. (1988), "On the Stratification of Skewed Populations," Survey Methodology, 14, 33-43.

Rao, J.N.K. (1991), "Variance Estimation under Imputation for Missing Data," unpublished manuscript.

Sanyal, S.K. and Sinha, S.K. (1977), "Methodological Problems in Large Scale Sample Surveys - Experiences from National Sample Survey," Sankya, Series c, 39, 47-70.

Särndal, C-E. (1990), "Methods for estimating the precision of survey estimates when imputation has been used," in Proceedings of the International Symposium on Measurement and Improvement of Data Quality, Statistics Canada.

Schiopu-Kratina, I. and Srinath K.P., (1991), "Sample Rotation and Estimation in the Survey of Employment, Payrolls and Hours," Survey Methodology, 17. No. 1.

Sethi, Y.K. (1963), "A Note on Optimum Stratification of Populations for Estimating the Population Means," Australian Journal of Statistics, 5, 20-33.

Srinath, K.P. (1987), "Methodological Problems in Designing Continuous Business Surveys: Some Canadian Experiences," Journal of Official Statistics, 3, 283-288.

Tambay, J.L. (1988), "Collocated Sampling," Statistics Canada, unpublished manuscript.