SAMPLING AND ESTIMATION FOR ESTABLISHMENT SURVEYS: STUMBLING BLOCKS AND PROGRESS

M.A. Hidiroglou, Statistics Canada
M.A. Hidiroglou, 11-J, R.H. Coats Bldg., Ottawa Canada, K1A 0T6

KEY WORDS: response burden, outliers, allocation

1. Introduction

Establishment surveys are repeated monthly, quarterly or annually to produce estimates of totals, averages, ratios, and changes between time periods for several characteristics of interest. The sampling frame for such surveys is highly dynamic. This causes several stumbling blocks both in the sampling and the estimation.

From the sampling point of view, changes in the frame mean that the stratification, sample size determination and allocation will be based on variables that are out of date. So, the resulting sample design is not efficient. What remedies do we have? For monthly surveys, because the frame is dynamic, this implies that it is difficult to maintain it with respect to classification changes and births. This also implies that the frame for a given reference month really refers to a "real world" frame that is older. How do we consider this?

Response burden is also another issue. This is specially so for large units that need to be included in the sample to obtain good estimates (and avoid outliers) of both level and change. It manifests itself in one of two ways. First, for a given survey, a sampled unit may be surveyed until it becomes out-of-scope to the survey. Establishments for whom time is money resent such lengthy and continuous response burden. Second, many surveys may request data from the same unit. This is especially the case for large units in take-all strata. How do we manage this burden?

From the estimation point of view, changes in the frame imply that domain estimation must be used to produce unbiased estimates. There is a point at which this estimation becomes ineffective. A re-stratification of the frame, new allocation of the sampling rates, as well as a redraw (partial or total) of the sample is required. How should one minimize the impact of a redraw on the estimates? Since business frames are highly skewed, there will be from time to time the problem of handling outliers. There are several questions that remain unanswered with respect to this problem. They are as follows: (a) At what level of aggregation should this detection and treatment occur? (b) How much discontinuity are we willing to accept to the published results between survey occasions, (c) How much bias is acceptable, given that present techniques for handling these units either decrease the sampling weight or trim the data values, (d) How do we handle outliers where exact linear relationships are obeyed by the data, but a data item is an outlier?

2. Sampling

2.1 Impact on Reliability

The first problem that we will discuss is the impact of an out-of-date frame on the reliability of estimates. Domain estimation must be used to produce unbiased estimates. In many business surveys, the domains of interest often coincide with the stratification. An example is industrial and geographical classification. Units are classified to a given stratum before sampling has occurred. After sampling, they may actually belong to another stratum. Suppose that the sample size has been computed with respect to the original classification on the frame. Changes in classification will disturb the levels of reliability sought for.

Suppose that we have L strata in the population U_h (h =1, 2, ..., L), which also happen to be our domains of interest (U_d : d = 1, 2, ... , L). Simple random samples s_h of size n_h are drawn from U_h, with sizes N_h, without replacement. With changes in classification, U_h can be decomposed into mutually exclusive and exhaustive sets $U_{h(d)}$ $(d=1,2,\ldots,L)$. When h = d, no change in classification has occurred. When

$h \neq d$, a change in classification has occurred.

The new population domains, $U_d^* = \bigcup_{h=1}^{L} U_{h(d)}$, (d=1, 2, ..., L), are no longer the same as the original population domains U_d, where, $U_d = \bigcup_{h=1}^{L} U_{d(h)}$. Suppose that for stratum d the level of precision for an estimated total was designed as "c" (the coefficient of variation). That is, $c = \sqrt{V(\hat{Y}_1(d))} / Y_1(d)$ with

$$Y_1(d) = \sum_{h=1}^{L} \Sigma_{U_{d(h)}} y_k = \Sigma_{U_d} y_k$$

and

$$\hat{Y}_1(d) = \sum_{h=1}^{L} \Sigma_{s_{d(h)}} \frac{N_d}{n_d} y_k(d) = \Sigma_{s_d} \frac{N_d}{n_d} y_k .$$

The variance fo $\hat{Y}_1(d)$ is

$$V(\hat{Y}_1(d)) = N_d^2 \left(1/n_d - 1/N_d\right) S_d^2$$

where $S_d^2 = \frac{1}{N_d - 1} \Sigma_{U_d} (y_k - \tilde{Y}(d))^2$

and $\tilde{Y}(d) = \Sigma_{U_d} y_k / N_d$.

The population total for a given new domain U_d^* , $Y_2(d) = \sum_{h=1}^{L} \Sigma_{U_{h(d)}} y_k$ will be estimated by $\hat{Y}_2(d) = \sum_{h=1}^{L} \frac{N_h}{n_h} \Sigma_{s_{h(d)}} y_k$

where $s_{h(d)}$ denotes the realized sample from $U_{h(d)}$.

We were originally targeting the precision for estimating $Y_1(d)$. With changes in classification, we are really estimating Y_2 (d). The resulting level of reliability for $\hat{Y}_2(d)$ may be significantly different from the one originally targeted. Note that $\Sigma_{U_d} (y_k - \tilde{Y}(d))^2$ can be written as :

$$\sum_{h=1}^{L} \Sigma_{U_{h(d)}} (y_k - \tilde{Y}(d))^2 + \sum_{h=1}^{L} \{(N_d(h) - 1) S_d^2(h) + N_d(h) (\tilde{Y}_d(h) - \tilde{Y}(d))^2\}$$

with $\tilde{Y}_d(h) = \Sigma_{U_{d(h)}} y_k / N_d(h)$,

$$S_d^2(h) = \frac{\Sigma_{U_{d(h)}} (y_k - \tilde{Y}_d(h))^2}{N_d(h) - 1} ,$$

and $N_d(h)$ is the number of population units belonging to $U_{d(h)}$.

The difference between $V(\hat{Y}_2(d))$ and $V(\hat{Y}_1(d))$ can be written as:

$$\sum_{h=1}^{L} \left\{ \frac{N_h^2}{N_h - 1} (1/n_h - 1/N_h) \left[(N_h(d) - 1) S_h^2(d) + N_h^2(d) \left(1/N_h(d) - 1/N_h\right) \tilde{Y}_h^2(d) \right] - N_d^2 \left(1/n_d - 1/N_d\right) \left[(N_d(h) - 1) S_d^2(h) + N_d(h) (\tilde{Y}_d(h) - \tilde{Y}(d))^2 \right] \frac{1}{N_d - 1} \right\}$$

which may be positive or negative, depending on the domain being considered.

2.2 Sample Size Determination

Sample sizes will be computed, bearing in mind that changes in classification occur. Since the stratification also happens to coincide with domains of interest, we would like to have the coefficient of variation "c" such that: $V(\hat{Y}_2(d)) = c^2 \hat{Y}_2^2(d)$.

Noting that

$$\Sigma_{U_h} (Y_k(d) - \bar{Y}_h(d))^2 = (N_h(d) - 1) S_h^2(d) + N_h^2(d) \left(1/N_h(d) - 1/N_h\right) \tilde{Y}_h^2(d)$$

we could estimate the mean and variance components from previous surveys. That is, we could suppose that $S^2(d) \doteq S_h^2(d)$, and that $\tilde{Y}(d) \doteq \tilde{Y}_h(d)$ $(h=1, 2, \ldots, L)$. The proportion of units expected to remain in domain U_d in stratum h, would be $p_h(d)$. The sample size for producing the estimate of a total for a given domain U_d would be:

$$n(d) = \frac{\sum_{h=1}^{L} N_h^2 p_h(d)\, b_h(d)/a_h}{c^2 \left[\sum_{h=1}^{L} N_h p_h(d)\, \tilde{Y}(d)\right]^2 + e(d)}$$

where

$$b_h(d) = S^2(d) + (1 - p_h(d))\, \tilde{Y}^2(d)$$

$$e(d) = \sum_{h=1}^{L} N_h p_h(d) \left[S^2(d) + (1 - p_h(d))\, \tilde{Y}^2(d)\right]$$

and a_h reflects the allocation scheme of the sample to the strata. The total sample size, n, could be the maximum of $n(d)$ $(d = 1, 2, \ldots, L)$ or a given quartile of the distribution of the $n(d)$'s.

2.3 Size Stratification

Efficient sampling of highly skewed populations such as those displayed by business surveys require that they be stratified into a take-all stratum and several take-some strata. The whole of units in the take-all stratum is selected with certainty, whereas units in the take-some strata are selected by a probability mechanism. Algorithms for stratifying a population into a take-all and take-some stratum have been given by Glasser (1962), Hidiroglou (1986), Lavallée and Hidiroglou (1988), and Hidiroglou and Srinath (1993). A problem with these schemes is that they do not consider the age of the size stratification variables.

For continuous surveys, such as monthly business surveys, the variable(s) used for size stratification are available regularly from the sample. The frame has an older version of this size stratification. If we model the newer version of the variable(s) on the older version, predicted values for the size stratification variable(s) can be obtained for all units on the frame. This implies that more up-to-date optimal boundaries and sampling rates can be computed. At a time of re-design the new sample can be drawn independently of the existing sample. However, we may wish to revise our stratum boundaries and sampling rates with an existing survey. This implies that the overlap between the new and old samples (Hoyt and Duggan, 1992) should be maximized. An overlap will ensure that the estimates between the two surveys will not change substantially. A model-based procedure is now suggested to carry this out.

Let y_k be the current variable of interest and $\boldsymbol{x}_k$ a p-dimensional vector of variables available for stratification on the older frame. Suppose that a regression model of the following form holds: $y_k = \boldsymbol{x}_k \beta + e_k$ with $E_\xi(e_k) = 0$, $E_\xi(e_k e_l) = 0$, $E_\xi(e_k^2) = \sigma_e^2$, k=1, 2, ..., N.

We assume simple random sampling in the take-some strata, and a regression model with an intercept term. The predicted value of the total $Y = \Sigma_U y_k$ (Särndal, Swensson and Wretman 1992) is

$$\hat{Y} = \Sigma_U \hat{y}_k + \Sigma_s w_k (y_k - \hat{y}_k) = \Sigma_U w_k \hat{y}_k ,$$

where $\hat{y}_k = \boldsymbol{x}_k \hat{\boldsymbol{B}}$ with

$$\hat{\boldsymbol{B}} = (\Sigma_s w_k \boldsymbol{x}_k \boldsymbol{x}'_k)^{-1} \Sigma_s w_k \boldsymbol{x}_k y_k ,$$ and w_k is the sampling weight.

Recalling that,

$$\Sigma_s (y_k - \bar{y}_s)^2 = \Sigma_s (y_k - \hat{y}_k)^2 + \Sigma_s (\hat{y}_k - \bar{y}_s)^2 ,$$

with $\bar{y}_s = \Sigma_s y_k / n$, we have $s_y^2 = s_e^2 + \hat{\boldsymbol{B}}' \boldsymbol{s}_{xx} \hat{\boldsymbol{B}}$. Here, s_e^2 is the sample variance of the y-values, and $\boldsymbol{s}_{xx}$ is the sample covariance matrix of the **x**'s. A corresponding expression for the population is $S_y^2 = S_e^2 + \boldsymbol{B}' \boldsymbol{S}_{xx} \boldsymbol{B}$, where $\boldsymbol{B} = (\Sigma_U \boldsymbol{x}_k \boldsymbol{x}_k')^{-1} \Sigma_U \boldsymbol{x}_k y_k$.

Now, under the model $E_\xi (s_e^2) = \frac{n-p}{n-1} \sigma_e^2$ (C.R. Rao, 1965), and $E_\xi (s_y^2) = \sigma_e^2 + \boldsymbol{B}' \boldsymbol{s}_{xx} \boldsymbol{B}$. This implies that

$$E_\xi (\hat{\boldsymbol{B}}' s_{xx} \hat{\boldsymbol{B}}) = E_\xi (s_y^2 - s_e^2) = \frac{n-p}{n-1} \sigma_e^2 + \boldsymbol{B}' s_{xx} \boldsymbol{B}$$

Hence, $\tilde{S}_y^2 = s_e^2 + \hat{\boldsymbol{B}}' \boldsymbol{s}_{xx} \hat{\boldsymbol{B}}$ can be used to estimate S_y^2. Since $\mathbf{S}_{xx}$, the population covariance matrix of the $\boldsymbol{x}$'s is available, a better estimator is $\hat{S}_y^2 = s_e^2 + \hat{\boldsymbol{B}}' \boldsymbol{S}_{xx} \hat{\boldsymbol{B}}$. This term can be used in the expressions for boundary determination, sample size determination and allocation.

We illustrate its use for a take-all stratum and a take-some stratum. Suppose that "c" is the required level of precision. Predict y_k as $\hat{y}_k = \boldsymbol{x}'_k \hat{\boldsymbol{B}}$ $(k = 1, 2, \ldots, N)$ and rank $\hat{y}_k$ from largest to smallest. Compute

$$n(t) = t + \frac{(N-t)^2 \hat{S}_y^2 (t)}{c^2 \hat{Y}^2 + (N-t) \hat{S}_y^2 (t)}$$

where $\hat{S}_y^2 (t) = s_e^2 + \hat{\boldsymbol{B}}' \boldsymbol{S}_{xx} (t) \hat{\boldsymbol{B}}$ denotes the estimate of the variance after having removed the largest t units. The optimum value for cut-off denoted as y^* is found when $n(m-1) \geq n(m)$ and $n(m) \leq n(m+1)$.

Extensions to more than one take-some stratum, can be obtained.

2.4 Frame and Sample Maintenance

Frame and sample maintenance consist of the following: (i) Updating the frame with new businesses (births) and sampling these births, so that the resulting sample represents the frame, (ii) Identifying businesses on the frame that are no longer in operation (deaths), (iii) Keeping track of classification changes on the frame and reflecting them in a representative manner in the sample, (iv) Reducing the response burden via sample rotation or partial replacement of the sample.

A problem with adding births to the frame is that there is a considerable time-lag between their "real world" start and the time they become available on the frame for sampling. Three ways to counter this effect are as follows. First, once they have been selected, impute the data they would have provided by applying inverse trends computed from the cell they belong to. Second, many of these units will be partially classified, but not available for sampling. Prorate the count of these units and add this count to the universe count, thereby inflating the sampling weight (Brown, Britney and Roumelis, 1991). Third, between sampling and estimation, a few survey cycles have gone by. An update to the population count can be obtained at the time of estimation and the weight for sampled units can be adjusted accordingly. For this, it is assumed that the sample still represents the more current frame.

Identification of businesses that are no longer operating takes place primarily through the

sample. Eliminating the inactive units from the sample without a corresponding elimination of units in the population, may lead to a bias in the estimates. This bias will occur, if the weights involve the known population and sample sizes. If such units are retained in the sample, the estimate is unbiased, but they contribute a zero component to the variance, thereby inflating it. A simple unbiased procedure to remove deaths from the frame and therefore the sample, is to use a source independent of the sampling process. Such a source may be difficult to identify if the frame is being updated through several sources. Annual surveys can also be used to update the frame. However, they usually lag the monthly surveys at least two years. This has a double negative impact. First, deaths are not current, since the sample will contain dead units that are at least two years old. Second, when the deaths are removed, their mass removal may cause blips in the estimates, depending on how the weight is computed. The impact on the estimates is minimized when the weighting procedure is a simple ratio of the number of units in the frame divided by the number of units in the sample. A procedure that gets around this problem is to remove dead units proportionately both in and out of sample. This keeps the weights stable, thereby avoiding artificial changes in the estimates of trends.

There are also changes to classification information. They include changes in industry, size and geography. Such changes, detected for in-sample units more often than for out-of-sample units, are handled via domain estimation. As noted in section 2.1, this type of estimation can become inefficient. If the size of in-sample units grows too much in relation to other units in the same stratum, we will be faced with problems of overestimation. We will deal with this problem in section 3. If an outside source, such as an annual survey, updates the classification on the frame, these changes can be reflected in the sample by tracking them as births and deaths. That is, units that no longer belong to the stratum in which they were originally selected are treated as deaths in that stratum and removed. They are allocated as births to the new stratum. Maximizing the sample overlap between these changes is desirable since it will reduce artificial blips in the estimates. The problem of classification becomes more difficult if there is no outside source that updates the frame on a universal basis. A total re-selection of the sample may be the only sensible solution for this case. Any other replacement scheme may introduce biases.

Response burden can be measured in terms of how many occasions a unit is kept in sample, or how many surveys are simultaneously contacting it. It can be minimized by sample rotation or partial replacement of the sample at each occasion. Schemes that allow rotation of the sample should allow the production of unbiased estimates for the parameter(s) of interest. Several schemes exist for carrying out rotation. Collocated sampling as described in Brewer, Early and Joyce (1972), first equispaces the population units on the [0,1) interval and then slightly disturbs the resulting number by adding a small random number. The scheme allows for the sample to be of fixed size or to have a fixed sampling fraction. Rotation takes place by sliding a "sampling window" in the interval [0,1). An advantage of this scheme is that the [0,1) interval can be partitioned into non-overlapping zones for different surveys. The rotation occurs within each zone. Collocated sampling is currently being used at the Australian Bureau of Statistics. Variations of this scheme have been recently developed by Cotton (1989), Ohlson (1993), and Srinath and Carpenter (1993). There are several trade offs to consider with such rotation schemes. These are: (i) the number of occasions that a unit will be kept in-sample and out-of-sample; (ii) the sampling fraction; (iii)the degree of sample overlap between rotations for year over year comparisons; and (iv) the costs associated with introducing a new unit into the sample.

Some units will inevitably be contacted by many surveys because of their large size. For these units, every effort should be made by survey organisations to co-ordinate the questionnaires between surveys, so as not to request the same information. Such co-ordination requires the existence of a questionnaire base. This base would contain: (i) a list of the questions for each survey, and (ii) a contact base which keeps an up-do-date list of units being surveyed. The resulting combination of this knowledge would result in the automatic creation of personalized questionnaires. This would ensure that response

burden is kept to a minimum.

There is a need to unify these processes and incorporate them into a system. Such an approach is being adopted at Statistics Canada. This system is known as the Generalized Sampling System (GSAM) at Statistics Canada. The system can stratify populations, compute sample sizes according to rules provided by users. The method of selection which includes, maintenance, and rotation is a variant on the collocated sampling procedure. A detailed description of its sampling algorithms is available in Srinath and Carpenter (1993). Generalization will remove a big stumbling block in sampling. It will allow for flexibility, standardization of procedures and system development cost savings.

3. Estimation

The term estimation will be used to describe several processes that occur once the data have been collected, captured and edited. They include imputation for missing data, and weighting. These processes are becoming more model-assisted, as the availability of auxiliary data can be used to improve the overall estimation. Weighting encompasses the use of auxiliary data for the following: small area estimation, adjustment for nonresponse, outlier treatment, composite estimation, and regression estimation. This unification of estimation is allowing the development of general estimation packages. These packages offer the user a variety of possible estimators, including estimators of variance.

Several advances are occurring on these fronts. We will restrict ourselves to regression estimation, imputation, and outlier treatment.

3.1 Regression

In the past, almost all business surveys at Statistics Canada (STC) used customized estimation systems. While this approach has provided the flexibility to meet specific requirements of each survey, many resources were spent in the development and maintenance of these systems. These systems were hard wired. They produced fixed outputs that were not easily changed. Any -changes in the outputs were costly and time--consuming to implement. The system maintenance costs were significant because of the acquisition and upgrading of different software and hardware products. There has been a constant need to train new system developers due to staff rotation on the projects.

The majority of business surveys have common features that can be shared in a system. It is with this thought in mind that at STC we are developing a Generalized Estimation System (GES). GES has and will continue to standardize development strategies and methodologies. Our current version of GES has been used extensively in our Agricultural and Business surveys. For these surveys, GES has cut down the development of an estimation system significantly, saving time, effort and resources.

Other estimation software packages have been developed elsewhere using different approaches to the methodology framework. These include LINWEIGHT (Bethlehem and Keller, 1987), PC-CARP (Schnell et al., 1988), SUDAAN (Shah et al., 1989). Building such systems implies that there must exist a unified theory of sampling that is easily amenable to programming and to further generalization. For GES, the generalization is dependent on recent developments in survey sampling which have unified many estimation concepts. For sampling, it is essentially the theory given in Rao (1979). Since auxiliary data play an important role in the estimation process, it has been included as another feature of GES. For this, we are using the sample-assisted approach as given in Särndal, Swensson and Wretman, 1992. We decided to adopt their framework based on the theory of the generalized regression estimator. This has allowed us to classify and use a large family of estimator functions through the specification of a general regression model. This theory has permitted the use of auxiliary information to improve on the efficiency of the estimators, while achieving consistency with the known auxiliary totals. The use of auxiliary data is particularly important because of its availability in many business surveys. We have characterized a generalized regression estimator through the concepts of model level, model groups, model auxiliary variables and model variance. This has provided us with a structure that includes many

traditional estimators, such as the combined and separate Horvitz-Thompson and ratio estimators, post-stratified estimators as well as more complex estimators such as raking ratio. More details are available in Lee, Hidiroglou and Estevao (1993).

3.2 Imputation

In most establishment surveys, non-responding units are followed up to improve the response rates. This follow-up is usually carried out by mail for the small to medium size non-responding units and by telephone for the larger units. Although this follow-up improves response rates, there will be, however, a group of non-responding units that may be classified as hard core non-respondents or late respondents. Hard core non-respondents are units that require much follow-up to respond, if at all. Late respondents are units that respond late with respect to the survey's collection cutoff date. The data for these units must therefore be imputed.

Units with no response will be called total non-respondents and those with partial non-response will be called partial non-respondents. The procedure used to impute the missing values depends on the availability of administrative files with auxiliary data. These auxiliary data should be well correlated with the variables to be imputed. For periodic establishment surveys, auxiliary data have been provided at some time by the non-responding unit. For monthly surveys, monthly ratios can be applied to last month's reported or imputed value. Annual ratios are used mostly for units that are seasonal that fail to provide a response as they emerge from their out-of-season period. These ratios are computed at a level (imputation cell) that usually corresponds to the original classification of these units. Imputation cells will be collapsed in a predetermined pattern if there are not enough units to compute the ratios. Dead or inactive units are imputed a value of zero. Births will either be imputed with the mean of imputation cell or using nearest neighbour.

Building an imputation system can be a labour intensive and time consuming task. Traditionally, the following four steps are involved: (i) analyze data to determine the most suitable imputation procedure, (ii) provide methodological specifications to reflect the imputation rules, (iii) build a custom made computer system, and (iv) test the system.. The availability of generalized imputation software can remove some of these steps. Such a system requires that the determination of imputation rules and their input into the system. Such a system has been developed at Statistics Canada: it is called the Generalized Edit and Imputation System (GEIS). GEIS offers the user a variety of commonly used imputation procedures that can be easily specified. Furthermore, the editing facilities in GEIS ensure that the imputed data obey expected linear relationships. This software has also removed a stumbling block.

Valid variance estimates for surveys with imputed data is not a trivial matter. It is well known that the standard variance estimators underestimate the true variance when applied to data with imputed values. Recently, Särndal (1990), and Rancourt, Särndal, and Lee (1994) have provided variance expressions for data imputed with the mean, ratio or nearest neighbour using model-assisted arguments. Rao and Shao (1992), and Kovar and Chen (1994), on the other hand have approached the problem using jackknife. These variance expressions apply to the Horvitz-Thompson estimator. Simulation studies have shown that these procedures produce good estimators of variance for data with imputed values. Generalized estimation systems such as GES, use auxiliary data via a linear regression model. No procedures exist presently to handle the variance of imputed data used as input into those models. Providing them would remove a further stumbling block in estimation.

3.3 Outlier Treatment

Outliers are units with legitimate values. It has been verified with the respondent that they are real. A few of these units in the sample will cause the estimates resulting from unbiased procedures to be unacceptable. There are two main design-based approaches used for the estimation of totals (or means) in the presence of outliers.

Once reduces the sampling weights of the outliers and suitably adjust the weights for the non-outliers (Hidiroglou and Srinath 1981; Ghangurde 1989). The other reduces (winsorises) the data value of the outliers (Searls 1966; Ernst 1980; Fuller 1991). When auxiliary variables are available, regression estimation can be used. Since these estimators are also outlier-prone, they must be made robust to outliers. Robust estimation as introduced by Huber (1973) has been adapted to finite population sampling by Chambers (1986). Lee (1991) has provided detailed review of these estimators.

The application of such procedures to periodic business surveys is not straightforward. Complicating factors include the dynamic nature of their samples. The sample composition changes due to births, deaths, and rotation. Furthermore, domain estimates (which reflect classification changes) may be required at several levels: industry, geographical (provinces or sub-province), and size. Treatment of outliers on a given survey occasion will have an impact on subsequent occasions. That is, we must also control the estimate of change. This implies that a "memory" of outlier treatment in previous occasions should be incorporated in the current estimation process.

Suppose that the estimator of total for a survey occasion "t" and domain d is $\hat{Y}^t(d) = \Sigma_{s^t} w_k^t a_k^t y_k^t(d)$ where $y_k^t(d)$ is the value of the k - h sampled unit at time t for domain U_d; w_k^t is the weight; and a_k^t is an adjustment factor (the memory) for outliers. The adjustment factor a_k^t is set to 1 initially, but it can change to reflect the dampening (via weight or data) of outlier units. The weight w_k^t can change through time, if it is computed as the ratio of the number of population units to the number of sampled units at time "t". It may also be adjusted across a number of strata to avoid accumulation bias. That is, if N_h^t is the number of populations units belonging to stratum h at time t, the weight for stratum h will be defined as $w_k^t = \frac{N_h^t}{n_h^t} \frac{\Sigma N_h^t}{\Sigma n_h^t w_h^t}$, where, n_h^t is the number of units in the sample, and w_h^t is the original sampling weight for stratum h at time t.

We suggest a procedure for dealing with outliers in business surveys. Following Tambay (1988) we decompose the estimate of change between two occasions t and s (t > s) as:

$$\Sigma_{s_c^t} w_k^t (y_k^t(d) - y_k^s(d)) + \Sigma_{s_c^t} (w_k^t - a_k^s w_k^s) y_k^s(d) + \Sigma_{s_I^t} w_k^t y_k^t(d) - \Sigma_{s_o^t} a_k^s w_k^s y_k^s(d)$$

where s_C^t refers to the common units in the sample, s_I^t represents units joining the sample as a result of births or rotation, and s_O^t represents units that have left the sample due to deaths or rotation. Two-sided outlier procedures are applied to the weighted differences in levels for units in the sample on two consecutive occasions. It should be noted that the impact of the term involving differences in weights between the occasions is negligible. One-sided detection procedures can be applied to units joining the sample. Two-sided detection procedures can also be applied to new units by suitably matching them against units that have left the sample. For units that have no corresponding match, a pseudo-difference can be created by subtracting the median of $a_k^s y_k^s(d)$, as obtained from the common units. Suitable adjustment factors a_k^t are then obtained. Fuller (1991)'s "test and treat" procedure can be used for this, treating positive and negative differences separately.

Any outlier treatment procedure will introduce downward bias into the estimate. Applying these procedures at low levels of aggregation will result in the accumulation of bias. These procedures should then be used at high levels of aggregation to avoid this.

References

Brewer, K.R.W., Early, L.J., and Joyce, S.F. (1972). Selecting Several Samples from a Single Population. *Australian Journal of Statistics*, 14, pp. 231-239.

Brown, A., Britney and Roumelis D.(1991). Adjustment of Estimates from Establishment Surveys for Undercoverage due to Frame Processing Lags. *Annual Research Conference,* Bureau of the Census, pp. 313-326.

Chambers, R.L. (1986). Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association*, 81, pp. 1063-1069.

Cotton, F. (1989) Use of SIRENE for Enterprise and Establishment Statistical Surveys. Paper presented at the 4th International Round Table on Business Frames.

Ernst, L.R. (1980). Comparison of Estimators of the Mean which Adjusts for Large Observations. *Sankhya Ser. C.*, 42, pp. 1-16.

Estevao, V., Hidiroglou, M.A. and Särndal, C.E. (1992). Requirements on a Generalized Estimation at Statistics Canada. Workshop on Uses of Auxiliary Information, Statistics Sweden .

Fuller, W.A. (1991). Simple Estimators of the Mean of Skewed Populations. *Statistics Sinica,* 1, pp. 137-158.

Ghangurde, P.D. (1989). Outliers in Sample Surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 736-739.

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, pp. 690-695.

Hidiroglou, M.A. (1986). The Construction of a Self-Representing Stratum of Large Units in Survey Design. *The American Statistician,* 40, pp. 27-31.

Hidiroglou, M.A., and Srinath, K.P. (1993). Problems Associated with Designing Subannual Business Surveys. *Journal of Business and Economic Statistics,* Vol.11, No. 4, pp. 397-405.

Hoyt, P., and Duggan, J. (1992). Recomputing Sample Sizes for an Ongoing Survey. *Proceedings of the Section on Business and economic Statistics,* American Statistical Association, pp. 203-208.

Huber, P.J. (1973). Robust Regression : Asymptotics, Conjectures and Monte Carlo. *Annals of Statistics*, 1, 799-821.

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

Kovar, J.G., and Chen, E.J. (1994). Jackknife Variance Estimation of Imputed Data. *Survey Methodology,* to appear.

Lee, H., Hidiroglou, M.A., and Estevao, V. (1993). GES: An Estimation System in Development at Statistics Canada.To appear in the *Proceedings of the Section on Business and Economic Statistics,* American Statistical Association.

Lee, H. (1991). Model-Based Estimator that are Robust to Outliers. *Proceedings of the 1991 Annual Research Conference,* U.S. Bureau of the Census, pp. 178-202.

Ohlson, E. (1993). Coordination of Samples Using Permanent Random Numbers. International Symposium on Establishment Surveys.

Rao, C.R. (1965) *Linear Statistical Inference and Its Applications*, New York, Wiley

Rao, J.N.K. (1979). On deriving mean square errors and their non-negative unbiased estimators in finite population sampling. *Journal of the Indian Statistical Association,* 17, pp. 125-136.

Rao, J.N.K., and Shao, J. (1992). Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. *Biometrika*, 79, pp. 811-822.

Rancourt, E., Särndal, C.E., and Lee, H. (1994). Estimation of variance in presence of nearest neighbour imputation. Paper presented the ASA meetings held in Toronto.

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

Särndal, C.E. (1990). Methods for Estimating the Precision of Survey Estimates when Imputation has been used. Proceedings of Statistic Canada 's Symposium'90: Measurement and Improvement of Data Quality, pp. 369-380.

Searls, D.T. (1966). An Estimator which Reduces Large Observations. *Journal of the American Statistical Association*, 61, pp. 1200-1204.

Srinath, K.P., and Carpenter, R. (1993). Sampling Methods for Repeated Business Surveys. International Symposium on Establishment Surveys.

Tambay, J.L. (1988). An Integrated Approach for the Treatment of Outliers in Sub-Annual Economic Surveys. *Proceedings of the Section Survey Research Methods,* American Statistical Association, pp. 229-234.