Social Statistics Research Centre
University of Southampton

Feeding back Information on Ineligibility from Sample Surveys to the Frame

Dan Hedlin, Suojin Wang

Abstract

It is usually discovered in the data collection phase of a survey that some units in the sample are ineligible even if the frame information has indicated otherwise. For example, in many business surveys a nonnegligible proportion of the sampled units will have ceased trading since the latest update of the frame. This information may be fed back to the frame and used in subsequent surveys, thereby making forthcoming samples more efficient by avoiding sampling nonnegligible units. We investigate what effect on survey estimation the process of feeding back information on ineligibility may have, and derive an expression for the bias that can occur as a result of feeding back. The focus is on estimation of the total using the common expansion estimator. We obtain an estimator that is nearly unbiased in the presence of feed back. This estimator relies on consistent estimates of the number of eligible and ineligible units in the population being available.

SSRC Methodology Working Paper M03/01

Feeding back information on ineligibility from sample surveys to the frame

DAN HEDLIN[1] and SUOJIN WANG[2]

ABSTRACT

It is usually discovered in the data collection phase of a survey that some units in the sample are ineligible even if the frame information has indicated otherwise. For example, in many business surveys a nonnegligible proportion of the sampled units will have ceased trading since the latest update of the frame. This information may be fed back to the frame and used in subsequent surveys, thereby making forthcoming samples more efficient by avoiding sampling nonnegligible units. We investigate what effect on survey estimation the process of feeding back information on ineligibility may have, and derive an expression for the bias that can occur as a result of feeding back. The focus is on estimation of the total using the common expansion estimator. We obtain an estimator that is nearly unbiased in the presence of feed back. This estimator relies on consistent estimates of the number of eligible and ineligible units in the population being available.

KEY WORDS: dead unit, feed back bias, overcoverage, permanent random number sampling, panel survey, co-ordinated samples.

[1] University of Southampton, Department of Social Statistics, Southampton SO17 1BJ, UK. e-mail: deh@socsci.soton.ac.uk

[2] Texas A&M University, Department of Statistics, College Station, Texas 77843-3143, USA. e-mail: sjwang@stat.tamu.edu

1. INTRODUCTION

To facilitate estimation of change, consecutive samples in a repeated survey are usually overlapping. If several surveys draw samples from the same frame, it is often desirable to spread the response burden out by making sure that samples for different surveys are not overlapping to a greater extent than necessary. This is particularly desirable if the frame is moderately large and used for many continuing surveys, which is a situation that many national statistical institutes face when conducting business surveys. Stratified simple random sampling is a very common design for business surveys. The skewed distribution of businesses calls for large sampling fractions in many strata, which aggravates the response burden for medium size and large businesses. Both estimation of change and response burden issues are of paramount importance in official business statistics. Therefore, sampling systems have been constructed that allow the organisation to co-ordinate samples, either positively or negatively (i.e. to create overlap or to make sure that there is little overlap).

For example, the Office for National Statistics (ONS) in the United Kingdom uses the Permanent Random Number (PRN) technique, which is a widely used method for drawing samples from lists. A PRN from the uniform distribution on [0,1] is attached to each frame unit independently of each other and independently of the unit labels and any variables associated with the units. Each unit will retain its PRN throughout its existence. The units can be plotted on a line starting at 0 and ending at 1 and we refer to this line as the *PRN line*. To draw a simple random sample without replacement, a *srswor*, with a

predetermined sample size *n*, a point is selected (randomly or purposively) on the PRN line and the *n* units to the right (say) are included in the sample. Two srswors are fully co-ordinated if they are drawn from the same interval. For overviews and further details see Ohlsson (1995) and Ernst, Valliant and Casady (2000). Table 1 shows starting points of sampling intervals of some of the business surveys the ONS conducts on a regular basis.

[Table 1 about here]

Samples for repeated surveys can also be selected with a panel technique where a set of rotation groups are selected at the first wave and one, say, of the groups is replaced with a fresh rotation group at the second wave and the other groups are retained in the sample. The difference between PRN sampling and panel sampling is more about the way to control overlaps than having different sampling designs.

There are in principle two main sources of data that are used to maintain a frame: administrative ones and surveys. Various administrative bodies send tapes to the ONS on a regular basis with information of, e.g., births and deaths of businesses. While these tapes are sent in to the ONS very frequently, the distribution of the time it takes for a new unit or an alteration of one old unit to come on to the frame is highly skewed. This is partly due to frame maintenance procedures, e.g. to avoid duplicates. There is also very often a considerable difference in time between the actual and formal termination of a business. Therefore, most of the ONS's business surveys share the information on deaths they obtain through their samples with other business surveys to speed up the information process. We examine the effects of using sample surveys to update a frame that is used

for repeated surveys. This is in principle how information of dead units is treated in business surveys at the ONS and some other national statistical institutes.

It would seem natural that this new information should be made available to other sample surveys, which otherwise may include the dead units in their samples and therefore lose precision. However, as pointed out by Srinath (1987) among others, such a procedure may cause bias. We refer to this as ***feed back bias***, which results whenever the sampling mechanism is not independent of the feed back procedure. For example, consider a situation where all dead units are found and deleted at the first wave of a panel survey. If no further deaths have occurred up to the second-wave observation of the panel units, the second-wave sample contains only live units. Without knowledge of the total number of live units in the population at the time of the second wave, an unbiased estimator of the total cannot be constructed. While more information about the population has been gathered when the deaths were recorded at the first wave, there is actually less information in the second wave-sample on the proportion of live units in the population.

A safe recommendation would be that no information on deaths from sample surveys, other than from completely enumerated strata, may be used to update the frame when samples are co-ordinated over time (cf. Ohlsson 1995, p. 168, and Colledge 1989, p. 103). However, to prohibit feeding back seems to deny oneself the use of all available information. We obtain an expression of the feed back bias and show that the feed back bias can be estimated and used to adjust conventional estimators. Schiopu-Kratina and Srinath (1991) adjust the sampling weights to counter an expected too low proportion of

dead units in the rotating sample of the Survey of Employment, Payroll and Hours conducted by Statistics Canada. Hidiroglou and Laniel (2001) discuss the feed back issue briefly. A general discussion of frame issues is given by Colledge (1995) and overviews of issues associated with continuing business surveys include College (1989), Hidiroglou and Srinath (1993), Srinath and Carpenter (1995), and Hidiroglou and Laniel (2001).

Instead of the terms eligible and ineligible we use the more emotive words dead and live, although our reasoning does cover all kinds of ineligibility. We confine our discussion to the estimation of the total of some study variable $\mathbf{y}' = (y_1, y_2, ..., y_N)$ on a population U with unit labels $\{1,2,...,N\}$,

$$t_y = \sum_U y_k \, . \qquad (1)$$

When the sampled units are observed, we assume that *all* dead units in the sample are classified as dead and the frame is updated with this information. This may be difficult in practice. In some surveys, however, the eligibility of all nonresponding units can be correctly identified.

Section 2 introduces the necessary notation and concepts and gives an expression for the feed back bias when estimating a total. Section 3 discusses three strategies that may be used in the presence of feed back and compares these in a simulation study. The paper concludes with a discussion in section 4.

2. AN EXPRESSION FOR FEED BACK BIAS

We assume throughout that a dead unit is always out of scope and that the value of the study variable of a dead unit is always zero. (It is conceivable that dead units are eligible in some surveys; for example, a business survey collecting data on production may have defined businesses that were alive at least a part of the reference period as eligible.) We adopt the design-based view that the survey population and the study variable are fixed and non-stochastic at any given point in time. The situation we address is as follows. One or more samples are drawn from the frame which comprises the *original survey population*, U_{orig}. For convenience we assume that the frame units and population units are of the same type. We refer to the updated frame, where all dead units that have been included in samples from U_{orig} have been excluded, as the *current survey population*, $U_{current}$. For example, two surveys may simultaneously work with a sample each, and after they have fed back, U_{orig} has shrunk to $U_{current}$. We disregard births of new units and other deaths than those deleted through samples from U_{orig}. We will also disregard undercoverage, nonresponse and measurement errors. In practice, administrative sources will provide information on deaths. They work independently from the sampling procedures employed by the statistical agency and will therefore not contribute to feed back bias. These units are *dead by administrative sources*. We can think of these dead units as being excluded from the population. While the sampling design is here assumed to be srswor, it can readily be extended to stratified simple random sampling.

Let U_d and U_l be the two subsets of the current survey population, $U_{current} = U_d \cup U_l$, that consist of dead and live units, respectively. A unit flagged on the frame as live belongs to either U_d and U_l. Units that are flagged as dead but for which the independence of detection and the sampling mechanism cannot be assured are called *dead by sample survey sources*. In our set-up, these are the dead units detected in samples taken from U_{orig}. Let the set of these units be denoted by U_{sd}, and we have the relationship $U_{orig} = U_{current} \cup U_{sd}$. Let N with a proper subscript be the size of each population, respectively. Then $N_{current} = N_l + N_d$, and $N_{orig} = N_l + N_d + N_{sd}$. At the time when samples are drawn from $U_{current}$, $N_{current}$ and N_{sd} are known numbers, whereas N_l and N_d are unknown. Moreover, N_{sd}, N_d and $N_{current}$ could be viewed as random depending on feed back results, while N_l is fixed. Following principles of Durbin (1969) and more recently in Thompson (1997), we would in many situations prefer to condition on N_{sd}. For example, if it is seen at the time when a sample is taken from $U_{current}$ that U_{sd} is in fact empty, then it does not seem appropriate to include in the inference the possibility that N_{sd} could have been large. However, to analyse the development of the feed back bias over a series of waves in a forthcoming panel survey, unconditional analysis would be preferable. We also provide an expression for the unconditional feed back bias.

[Figure 1 about here]

Denote by $U_{nodeads}$ the part of $U_{current}$ that was covered by the previous sample(s) drawn from U_{orig}; see Figure 1. Clearly, $U_{nodeads}$ is a random set depending on previous samples. Since $U_{nodeads}$ is winnowed from dead units we have $U_{nodeads} \subset U_l$. The complement to $U_{nodeads}$, denoted by $U_{withdeads}$, is also a random set and encompasses all of U_d and a part

of U_l. We have $U_{nodeads} \cup U_{withdeads} = U_l \cup U_d = U_{current}$. To derive the feed-back bias we will consider a sample of size n with a sample part s_a of size n_a taken from $U_{nodeads}$ through PRN sampling or a panel sampling technique, and the remaining part s_b is taken from $U_{withdeads}$. Let $I(k \in s_a) = 1$ when unit k is included in s_a, otherwise $I(k \in s_a) = 0$. Recall that $y_k = 0$ if k is a dead unit. Thus we have $\sum_{s_a} y_k = \sum_{U_l} y_k I(k \in s_a) = \sum_{U_{current}} y_k I(k \in s_a)$ and, assuming that $N_l > 0$, $\Pr[k \in s_a \mid k \text{ alive}, N_{sd}] = \frac{n_a}{N_l}$. The probability is conditional on unit k being alive since it is determined by design that only live units can be included in $U_{nodeads}$. Denote the bias of an estimator $\hat{q}$ for the parameter q by $B(\hat{q}, q)$. Then with respect to the population total $t_y = \Sigma_{U_{current}} y_k$, the bias of a general linear estimator $\hat{t}_y^{(s_a)} = \sum_{s_a} w_k y_k$, with any given w_k's, is

$$B(\hat{t}_y^{(s_a)}, t_y \mid N_{sd}) = \sum_{U_l} \{w_k E[k \in s_a \mid k \text{ alive}, N_{sd}] - 1\} y_k = \Sigma_{U_l} \left(\frac{w_k n_a}{N_l} - 1 \right) y_k$$

$$= \sum_{U_{current}} \left(\frac{w_k n_a}{N_l} - 1 \right) y_k \ . \qquad (2)$$

In particular, the bias of the expansion estimator $\hat{t}_y^{(s_a)} = \frac{N_{current}}{n_a} \sum_{s_a} y_k$ is

$$B(\hat{t}_y^{(s_a)}, t_y \mid N_{sd}) = \frac{N_d}{N_l} t_y \ . \qquad (3)$$

Alternatively, sampling of s_a can be seen as a two-phase sampling scheme. Note that in the first phase,

$$\Pr[k \in U_{nodeads} \mid k \text{ alive}, N_{sd}] = \Pr(k \in U_{nodeads}, k \text{ alive} \mid N_{sd}) / \Pr(k \text{ alive} \mid N_{sd})$$

$$= \frac{N_{nodeads}}{N_{orig}} \Big/ \frac{N_l}{N_{orig}} = \frac{N_{nodeads}}{N_l} \ . \qquad (4)$$

Thus,

$$\Pr\left[k \in s_a \middle| \ k \text{ alive}, N_{sd}\right] = \frac{N_{nodeads}}{N_l} \frac{n_a}{N_{nodeads}} = \frac{n_a}{N_l}. \qquad (5)$$

Note that $N_{nodeads}$ (and thus N_{sd}) cancels out. The probability of $(k \in s_a)$ depends on the feed back process to have taken place but not on the size of U_{sd}.

Next, to derive the bias for the sample part s_b of size n_b taken from $U_{withdeads}$, first note that $(k \in U_{withdeads})$ is the same event as $(k \notin U_p)$, where $U_p = U_{nodeads} \cup U_{sd}$ is the part of U_{orig} covered by previous samples. Then

$$\Pr\left[k \in U_{withdeads} \middle| N_{sd}\right] = \Pr\left[k \notin U_p \middle| N_{sd}\right] = \frac{N_{orig} - N_p}{N_{orig}} = \frac{N_{withdeads}}{N_{orig}} \ . \qquad (6)$$

This conditional probability again does not depend on the relative sizes of $U_{nodeads}$ and U_{sd}. On the other hand, the probability of including a unit in s_b given that feed back has occurred is

$$\Pr\left[k \in s_b \middle| N_{sd}\right] = \frac{n_b}{N_{withdeads}}. \qquad (7)$$

From (7) we obtain that the conditional expected value of $\hat{t}_y^{(s_b)} = \sum_{s_b} w_k y_k$ is

$$E\left(\hat{t}_y^{(s_b)} \middle| N_{sd}\right) = E\left[\frac{n_b}{N_{withdeads}} \sum_{U_{withdeads}} w_k y_k \ \middle| N_{sd}\right]$$

$$= \frac{n_b}{N_{withdeads}} \ \frac{N_l - N_{nodeads}}{N_l} \sum_{U_{orig}} w_k y_k \ .$$

The second equation above is due to the fact that given N_{sd}, all N_l live units in U_{orig} are equally likely to be in $U_{withdeads}$, which has $N_l - N_{nodeads}$ live units. Therefore, the conditional bias of $\hat{t}_y^{(s_b)}$ is

$$B\left(\hat{t}_y^{(s_b)}, t_y \middle| N_{sd}\right) = \sum_{U_{orig}} \left(\frac{w_k n_b}{N_{withdeads}} \frac{N_l - N_{nodeads}}{N_l} - 1 \right) y_k$$

$$= \sum_{U_{current}} \left(\frac{w_k n_b}{N_{withdeads}} \frac{N_l - N_{nodeads}}{N_l} - 1 \right) y_k \,. \qquad (8)$$

For the expansion estimator $\hat{t}_y^{(s_b)}$ with weights $w_k = N_{current}/n_b$ the bias is

$$B\left(\hat{t}_y^{(s_b)}, t_y \middle| N_{sd}\right) = B t_y, \qquad (9)$$

where

$$B = \frac{N_{current}}{N_{withdeads}} \frac{N_l - N_{nodeads}}{N_l} - 1 = \frac{N_{current}\left(N_l - N_{nodeads}\right) - N_l\left(N_{current} - N_{nodeads}\right)}{N_{withdeads} N_l}$$

$$= -\frac{N_d N_{nodeads}}{N_l N_{withdeads}} = -\frac{N_d \left(N_p - N_{sd}\right)}{N_l \left(N_{orig} - N_p\right)} \,.$$

The bias is always non-positive since $B \le 0$. It is easy to see that B is an increasing function of N_{sd} since $N_d = N_{totaldeads} - N_{sd}$, where $N_{totaldeads}$ is the fixed number of all dead units in U_{orig}. It is also readily seen that the maximum of B is attained when U_{sd} encompasses all dead units in U_{orig}, that is, when $N_{sd} = N_{totaldeads}$.

Combining (9) with (3) we obtain the overall bias of $\hat{t}_y = \frac{N_{current}}{n} \Sigma_{s_{current}} y_k$ to be

$$B\left(\hat{t}_y, t_y \middle| N_{sd}\right) = E(\hat{t}_y \mid N_{sd}) - t_y = \frac{N_d}{N_l} \left(\frac{n_a}{n} - \frac{n_b}{n} \frac{N_{nodeads}}{N_{withdeads}} \right) t_y = \tilde{c}\, t_y \,. \qquad (10)$$

The bias in the expansion estimator is really down to not knowing the correct population size. In (3) the bias stems from multiplying the sample average over live units with $N_{current}$ rather than the unknown N_l. The bias from the sample parts s_a and s_b will in absolute terms be less than (3) and (9), respectively, if some of the dead units in the samples from U_{orig} have not been identified as dead and therefore have not been weeded out. This would happen, for example, if the status of nonresponding units is difficult to determine.

An unconditional analysis in the presence of feed back can be obtained directly by taking expectation of (10) with respect to N_{sd}. Thus, unconditionally, we have

$$E\left(\frac{N_{current}}{n}\sum\nolimits_{s_{current}} y_k\right) - t_y$$

$$= \left\{\frac{N_{totaldeads} - E(N_{sd})}{N_l}\left(\frac{n_a}{n} - \frac{n_b}{n}\frac{N_p - E(N_{sd})}{N_{withdeads}}\right) - \frac{n_b}{nN_l N_{withdeads}}\right\} t_y = c\,t_y, \qquad (11)$$

where $E(N_{sd}) = N_p\, N_{totaldeads} / N_{orig}$.

Lavallée (1996) took an interesting approach to a similar problem with panel survey data. In that paper, the problem of frame update using panel with rotation is addressed among other issues. Our approach is different from the approach of that paper in that we consider the two conditional probabilities $\Pr[k \in s_a \mid k \text{ alive}, N_{sd}]$ and $\Pr[k \in s_b \mid N_{sd}]$ separately.

3. THREE SIMPLE STRATEGIES AND A SIMULATION STUDY

A strategy, which is referred to as Strategy 1 here, is to feed back, delete the set U_{sd} from the frame and accept the feed back bias. However, the size of the bias is seldom known. The estimator for Strategy 1 is $\hat{t}_y = \frac{N_{current}}{n} \sum_{s_{current}} y_k$ where $s_{current}$ is a sample taken from $U_{current}$. To obtain Strategy 2, note that if consistent estimates of N_d and N_l are available these may be plugged into (10) or (11) and an estimator with favourable properties is obtained:

$$\hat{t}_y' = \hat{t}_y (1+\hat{c})^{-1}, \qquad (12)$$

where $\hat{c} = \frac{\hat{N}_d}{\hat{N}_l}\left(\frac{n_a}{n} - \frac{n_b}{n}\frac{N_p - N_{sd}}{N_{orig} - N_p}\right)$ for both the conditional and unconditional cases since the term $n_b (n N_l N_{withdeads})^{-1}$ in (11) is negligible. The estimates $\hat{N}_d$ and $\hat{N}_l$ of the sizes of the domains U_d and U_l can be obtained from a sample from the original or current survey population with

$$y_k = \begin{cases} 1, \text{ if unit } k \in N_d \ (N_l), \\ 0, \text{ otherwise.} \end{cases}$$

As the following argument shows, we do not expect the bias of (12) to be large:

$$E\left(t_y'\right) = E\left[\hat{t}_y (1+\hat{c})^{-1}\right] \approx E\left(\hat{t}_y\right)(1+c)^{-1} = t_y (1+c)(1+c)^{-1} = t_y .$$

Another strategy, here denoted by Strategy 3, is to feed back the information that certain units are dead, but to retain them on the frame and allow them to be sampled. In theory, the resulting estimator is unbiased, but the disadvantage of this strategy is that the

precision will suffer as part of the sample is lost on ineligible units. The estimator of Strategy 3 is $\hat{t}_y'' = \frac{N_{orig}}{n} \sum_{s_{orig}} y_k$.

A simulation study may shed some light on which of the Strategies 1-3 is to be preferred. Natural measures for comparing the strategies are bias and variance. In business surveys, estimates for subpopulations (industries) are often more interesting than the whole population. To simulate a subpopulation, a frame consisting of 1000 units was created to form the original survey population. A gamma distributed value, Y1, was associated with each unit. We used the same gamma distribution as the one that generated Population 12 in Lee, Rancourt, and Särndal (1994, p. 236). The coefficient of variation (population standard deviation divided by the mean) was 0.57. Another study variable, Y2, was created by performing independent Bernoulli trials, one for each population unit, which obtained value 1 with probability equal to 0.5 and value 0 otherwise. Unlike in Lee, et al., some of the units were dead. Each unit was independently of other units classified as dead with a probability P_{dead}. All dead units were assigned zero values for both Y1 and Y2. A set of Y1 and Y2 were simulated for each of four values of P_{dead}: 0.03, 0.05, 0.2, and 0.5. These sets contained 29, 54, 201 and 494 dead units, respectively.

A PRN was attached to each unit and the units were laid out along a PRN line. The first sample, s_1, was drawn by identifying the 500 units with the smallest PRNs. All dead units in s_1 were flagged as 'dead by sample survey sources'. Hence, U_p covered approximately the first half of the PRN line. The frame with the units flagged as dead by sample survey sources excluded made up the current survey population. The estimates of N_d and N_l used

in Strategy 2 were based on s_1. A second sample, denoted by $s_{2current}$, was drawn by taking 100 units to the right of a starting point, *start* 2, disregarding units dead by sample survey sources. Another sample of 100 units was selected from *start* 2, but units dead by sample survey sources were this time allowed to be included in this sample. Hence, this sample was drawn from U_{orig}, and we denote it by s_{2orig}. Figure 2 shows the PRN intervals and the study variable Y1.

[Figure 2 about here]

The procedure described in the preceding paragraph was repeated 1000 times. That is, for each of the values of P_{dead} mentioned above and for each of three starting points of s_2, to be defined, 1000 sets of PRNs were generated and attached to the units. The frame was reordered for each new set of PRNs, and three samples were drawn for each reordering (s_1, $s_{2current}$, and s_{2orig}). Two values of *start* 2, 0.0 and 0.7, were chosen so as to make the proportion of $s_{2current}$ that fell in U_{nodead} 100% and 0%, respectively. That is, n_a/n was set to 100% and 0%. Further, to make n_a/n on average 50% under each of the chosen P_{dead}, appropriate values of *start* 2 were derived. They are 0.448, 0.447, 0.438, and 0.4 for the P_{dead} values 0.03, 0.05, 0.2, and 0.5, respectively.

In summary, the population and samples sizes, the study variables Y1 and Y2, and which of the units that were dead were held fixed in our study. For twelve combinations of P_{dead} and n_a/n, the reordering of the units on the PRN line through the simulation of new PRNs made the following factors vary:

- which of the units that were included in s_1, $s_{2current}$, and s_{2orig};
- how many and which of the dead units that were dead by sample survey sources;

- which of the units that belonged to $U_{nodeads}$ and $U_{withdeads}$.

Thus the quantities N_{sd}, N_d and $N_{current}$ vary in the simulations. It seems practical to let them do so rather than to control them in an experiment with more factors than P_{dead} and n_a/n.

Table 2 shows the empirical relative bias of Strategies 1 and 2, computed as the straight average of the 1000 differences between the estimate and the parameter in terms of the percentage of the total obtained in the simulation. Strategy 3 is unbiased and is therefore not included in Table 2. The bias of Strategy 3 that nevertheless appeared in the simulations reflects the simulation error; it was at most 0.5%. As seen in Table 2, Strategy 2 is virtually unbiased as well. Note that the simulated bias under Strategy 1 is what (11) predicts (with allowance for simulation error). This bias is appreciable in nearly all cases and if the proportion of dead (or ineligible) units is high the bias can be very severe indeed. Table 3 shows the empirical coverage probabilities. While Strategy 2 gives in all cells coverage probabilities close to the targeted 95%, Strategy 1 achieves that in general only for the population with 3% dead units. The coverage probability under Strategy 1 tends also to be acceptable for populations with a larger proportion of dead units, if half of the sample is taken from the part of the PRN line where dead units have been weeded out, and the other half from the part of the PRN line where the original proportion of dead units has been retained, as the negative bias from the first half of the sample tends to cancel out the positive bias from the second half.

[Table 2 about here]

[Table 3 about here]

The variance of the simulated estimates was computed. Tables 4 and 5 show the variance of Y1 and Y2, respectively, under Strategies 2 and 3 relative to that of Strategy 1, which in all cases gives a smaller variance than Strategy 3. Hence, considering the extra complexity of Strategy 2, the feed back strategy seems preferable for populations with a small proportion of ineligible units, say 3% or less. If this proportion is larger than, say, 5%, the bias of Strategy 1 may cause poor coverage probabilities and misleading estimates. The variance of Strategy 2 is no worse than that of Strategy 3; in most cases Strategy 2 is superior. The non-monotone variance ratios in the bottom row of Table 4 is due to the estimation of N_d and N_l combined with the specific details of the simulation.

[Table 4 about here]

[Table 5 about here]

4. DISCUSSION

This paper gives conditional and unconditional expressions for the feed back bias when the total is estimated with the common expansion estimator. We have shown that the feed back bias can be large. With as little as 5% ineligible units on the frame, feeding back information of these from sample surveys can result in about 2-3% bias. However, a small-scale simulation study indicates that if the proportion of ineligible units is 3% or less, the feed back strategy does not seem to create problems in terms of bias and variance.

We have also derived a virtually unbiased estimator. The simulation study shows that this estimator compares favourably in terms of variance with the alternative strategy of

retaining ineligible unit on the frame and letting them be included in further samples. This estimator relies on the availability of consistent estimates of the number of eligible and ineligible units in the population. These estimates may be obtained from an earlier sample in which the unbiased strategy of letting units that have been found dead be included in the sample.

In order to facilitate the theoretical development, we have made simplifying assumptions. The most important of these is the assumption that *all* dead units have been found in earlier sample surveys and have been fed back to the frame. We have envisaged a frame with one 'white' area, where all ineligibles have been flagged as such, and one 'black' area, where no ineligibles have been touched. In practice, this is not likely to happen. If the frame is moderately large and used for many continuing surveys, some of which may feed back to varying intensity, the frame will turn 'grey' rather than 'black and white'. Clearly, the feed back bias will then be less severe than in the 'black and white' situation. It has not, however, been in the scope of this paper to quantify the bias for a 'realistically grey' frame. In this sense, what has been examined in this paper is a worst case scenario.

ACKNOWLEDGMENT

The authors thank Mark Pont for very useful initial discussions of this topic and two referees for very valuable comments. Both authors' research was partially supported by the U.K. Office for National Statistics and Wang's research was also supported by the U.S. National Cancer Institute (CA 57030).

REFERENCES

COLLEDGE, M.J. (1989). Coverage and Classification Maintenance Issues in Economic Surveys. In *Panel Surveys*. (Eds. Kasprzyk, D., Duncan, G. J., Kalton, G., and Singh, M. P.). New York: Wiley, 80-107.

COLLEDGE, M.J. (1995). Frames and Business Registers: An Overview. In *Business Survey Methods*. (Eds. Cox, B., Binder, D., Chinappa, N., Christianson, A., Colledge, M. and Kott, P). New York: Wiley, 21-47.

DURBIN, J. (1969). Inferential Aspects of the Randomness of Sample Size in Survey Sampling. In *New Developments in Survey Sampling.* (Eds. Johnson, N.L. and Smith, H.). New York: Wiley, 629-651.

ERNST, L.R., VALLIANT, R., and CASADY, R.J. (2000). Permanent and Collocated Random Number Sampling and the Coverage of Births and Deaths. *Journal of Official Statistics*, 16, 211-228.

HIDIROGLOU, M.A. and LANIEL, N. (2001). Sampling and Estimation Issues for Annual and Sub-Annual Canadian Business Surveys. *International Statistical Review*. 69, 487-504.

HIDIROGLOU, M.A. and SRINATH, K.P. (1993). Problems Associated with Designing Subannual Business Surveys. *Journal of Business and Economic Statistics*, 11, 397-406.

LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1994). Experiments with Variance Estimation from Survey Data with Imputed Values. *Journal of Official Statistics*, 10, 231-243.

LAVALLÉE, P. (1996). Frame Update Problems with Panel Surveys. *Proceedings of Statistical Days '96*, Statistical Society of Slovenia, 252-261.

OHLSSON, E. (1995). Coordination of Samples Using Permanent Random Numbers. In *Business Survey Methods*. (Eds. Cox, B., Binder, D., Chinappa, N., Christianson, A., Colledge, M. and Kott, P). New York: Wiley, 153-169.

SCHIOPU-KRATINA, I. and SRINATH, K.P. (1991). Sample Rotation and Estimation in the Survey of Employment, Payrolls and Hours. *Survey Methodology*, 17, 79-90.

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

SRINATH, K.P. and CARPENTER, R.M. (1995). Sampling Methods for Repeated Business Surveys. In *Business Survey Methods.* (Eds. Cox, B., Binder, D., Chinappa, N., Christianson, A., Colledge, M. and Kott, P). New York: Wiley, 171-183.

THOMPSON, M.E. (1997). Theory of Sample Surveys. London: Chapman & Hall.

Table 1. Starting points of the PRN sampling intervals of some of the business surveys the UK Office for National Statistics conducts

Survey	Starting point of sampling interval
The Monthly Inquiry for the Distribution and Services Sector, and other monthly surveys covering other sectors of the business population	0
The Quarterly Capital Expenditure Inquiry	0.125
The UK Survey of Products of the European Community	0.375
The Inquiry of Stocks	0.5
The Annual Business Inquiry	0.625
The Annual Employment Survey	0.75

Table 2. Bias, % of total of Y1. The first entry in each cell is the bias under Strategy 1, the second is the bias under Strategy 2

	Average of n_a/n					
P_{dead}	**0%**		**50%**		**100%**	
0.03	-1.6	-0.1	0.4	0.4	1.5	0.0
0.05	-2.8	0.0	0.4	0.4	2.9	0.0
0.20	-10.2	-0.2	1.5	0.4	12.7	0.1
0.50	-24.6	0.2	12.5	0.3	49.0	0.2

Table 3. The coverage probability in percentage for estimating total of Y1. The first entry in each cell is the under Strategy 1, the second is the coverage probability under Strategy 2.

	Average of n_a/n					
P_{dead}	**0%**		**50%**		**100%**	
0.03	94.6	94.3	94.6	94.8	94.3	95.1
0.05	93.3	95.2	94.4	93.9	90.8	95.0
0.20	65.9	94.5	93.8	94.8	46.1	94.6
0.50	21.2	95.1	78.4	94.7	0.0	94.8

Table 4. Variance ratio of the estimator of the total of Y1. The first entry in each cell is the variance under Strategy 2 relative to that of Strategy 1, the second is the variance under Strategy 3 relative to Strategy 1.

	Average of n_a/n					
P_{dead}	**0%**		**50%**		**100%**	
0.03	1.04	1.04	1.00	1.06	0.98	1.08
0.05	1.08	1.08	0.98	1.14	0.95	1.15
0.20	1.28	1.28	0.85	1.27	0.83	1.46
0.50	1.85	1.85	0.52	1.34	0.58	2.24

Table 5. Variance ratio of the estimator of the total of Y2. The first entry in each cell is the variance under Strategy 2 relative to that of Strategy 1, the second is the variance under Strategy 3 relative to Strategy 1.

	Average of n_a/n					
P_{dead}	**0%**		**50%**		**100%**	
0.03	1.03	1.03	1.00	1.03	0.97	1.03
0.05	1.06	1.06	0.99	1.04	0.95	1.06
0.20	1.25	1.25	0.92	1.15	0.80	1.19
0.50	1.80	1.81	0.65	1.40	0.50	1.36

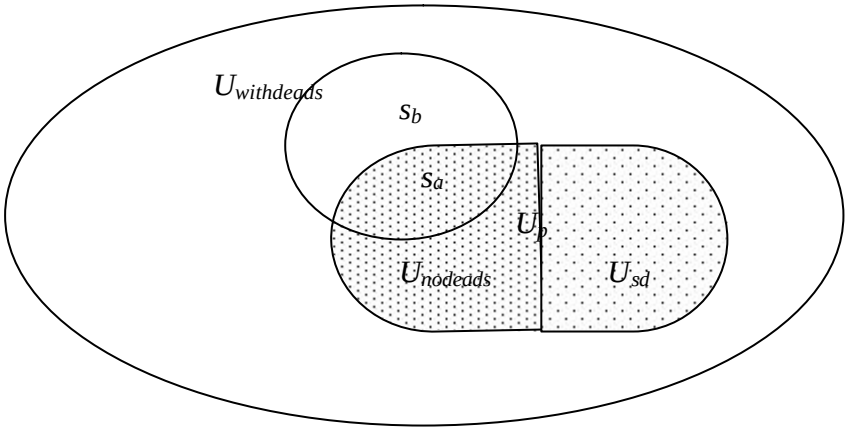

Figure 1. The original survey population, U_{orig}, and its subsets.

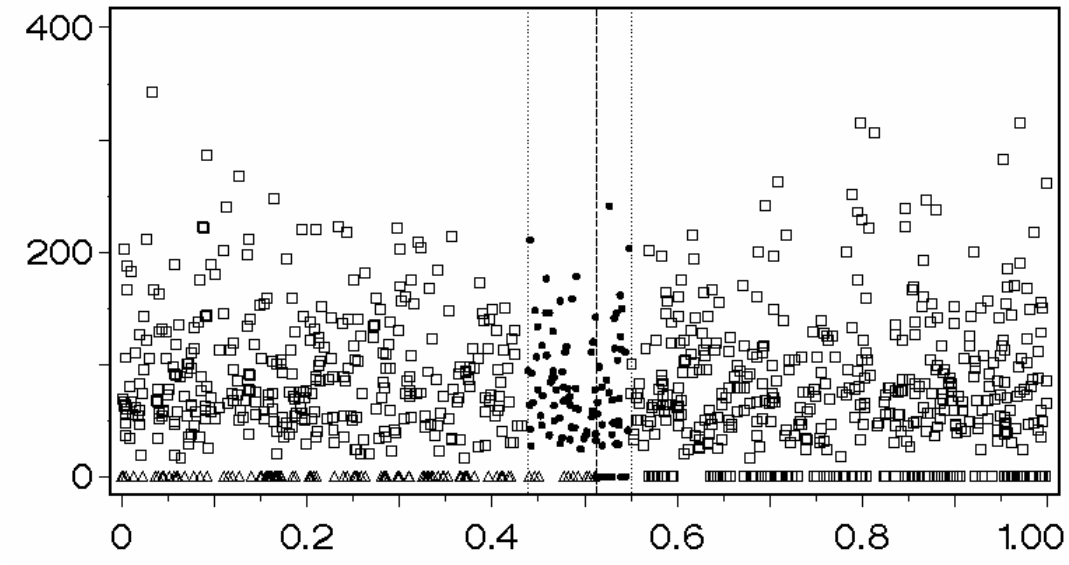

Figure 2. A plot of one of the simulated populations, the study variable Y1 against the PRNs, with P_{dead} = 0.20. The dots are units included in $s_{2current}$ (the sample from the current survey population); the triangles are units that are dead by statistical sources and squares represent units belonging to the current survey population but are not included in the sample from this population. The PRN interval for s_1 (the 500 units in the first sample from the original survey population) is (0, 0.51) and the one for $s_{2current}$ is (0.44, 0.55).