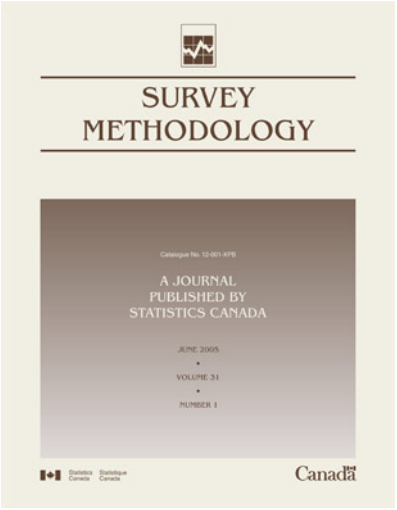
Catalogue no. 12-001-XIE

Survey Methodology

December 2004



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2004

Published by authority of the Minister responsible for Statistics Canada

April 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 30, No. 2, pp. 167-174
Statistics Canada, Catalogue No. 12-001

Feeding Back Information on Ineligibility from Sample Surveys to the Frame

Dan Hedlin and Suojin Wang [1]

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 ineligible units. On the first of two survey occasions, we assume that all ineligible units in the sample (or set of samples) are detected and excluded from the frame. On the second occasion, a subsample of the eligible part is observed again. The subsample may be augmented with a fresh sample that will contain both eligible and ineligible 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. An estimator that is nearly unbiased in the presence of feed back is obtained. 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; Coordinated samples.

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 the PRN throughout its existence. The units can be ordered along 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, an *SI*, 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 SIs 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).

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 on, *e.g.*, births and deaths of businesses. While these tapes are sent to the ONS very frequently, the distribution of the time it takes for a new unit or an alteration of an old unit to be registered on 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

1. Dan Hedlin, Statistics Sweden, Box 24 300, SE-104 51 Stockholm, Sweden. E-mail: dan.hedlin@scb.se; Suojin Wang, Texas A&M University, Department of Statistics, College Station, Texas 77843-3143, U.S.A. E-mail: sjwang@stat.tamu.edu.

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 on dead units is treated in business surveys at the ONS, Statistics Sweden, 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. We show how an estimate of the number of live units in the population can be used to construct an approximately unbiased estimate of the population total.

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, page 168, and Colledge 1989, page 103). However, to prohibit feeding back seems to deny oneself the use of all available information. We obtain an expression for 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. The discussion is confined to the estimation of the total

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

of some study variable $\mathbf{y}' = (y_1, y_2, ..., y_N)$ on a population U with unit labels $\{1, 2, ..., N\}$.

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 expressions 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. Expressions for Feed Back Bias

2.1 Introduction and Notation

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 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_1. Let the set of samples drawn from U_1 be denoted by s_1. 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_1 have been excluded, as the *current survey population*, U_2. For example, two surveys may simultaneously work with a sample each, and after they have fed back, U_1 has shrunk to U_2. We disregard births of new units and other deaths than those deleted through samples from U_1. 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. See Hidiroglou and Laniel (2001) for a discussion of estimation in the presence of units deathed by administrative sources. While the sampling design here is assumed to be SI, it can readily be extended to stratified simple random sampling.

Let $U_{2,d}$ and $U_{2,l}$ be the two subsets of the current survey population, $U_2 = U_{2,d} \cup U_{2,l}$, that consist of dead and live units, respectively. All units in $U_{2,d}$ and $U_{2,l}$ are assumed to be flagged as live. 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_1. Let the set of these units be denoted by $s_{1,d}$, and we have the relationship $U_1 = U_2 \cup s_{1,d}$. Figure 1 displays the sets and their relationships. Let N and n with a proper subscript be the size of the corresponding population and sample(s), respectively. Then $N_1 = N_2 + n_{1,d}$ and $N_2 = N_{2,l} + N_{2,d}$. At the time when samples are drawn from U_2, N_2 and $n_{1,d}$ are known numbers, whereas $N_{2,l}$ and $N_{2,d}$ are unknown. Moreover, $n_{1,d}$, $N_{2,d}$ and N_2 could be viewed as random depending on feed back results, while $N_{2,l}$ is fixed. Following principles of Durbin (1969) and more recently in Thompson (1997), we would in many situations prefer to condition on $n_{1,d}$. For example, if it is seen that $n_{1,d} = 0$, then it does not seem appropriate to include in the inference the possibility that $n_{1,d}$ could have been large. However, to analyse the development of the feed back bias over a series of waves in a panel survey when planning the survey, unconditional analysis would be preferable. We also provide an expression for the unconditional feed back bias.

Denote by $s_{1,l}$ the live part of s_1, *i.e.*, the part of U_2 that was covered by the previous sample(s) drawn from U_1; see Figure 1. Clearly, $s_{1,l}$ is a random set and we have $s_{1,l} \subset U_{2,l}$. Let the nonsampled part of U_2 be denoted by $U_{2,wd}$ ('wd' for 'with dead units'). It is also a random set and encompasses all of $U_{2,d}$ and part of $U_{2,l}$. We have $U_2 = U_{2,wd} \cup s_{1,l}$.

Let s_2 be an SI taken from U_2. Estimators based on s_2 will suffer from feed back bias unless special information is at hand, such as knowledge about $N_{2,l}$, which is not usually the case. To derive an expression for the feed back bias we shall first obtain the inclusion probabilities. To do this, it is useful to consider the two sample parts of s_2 separately: the sample part $s_{2,a}$ of size $n_{2,a}$ taken from $s_{1,l}$ through PRN sampling or a panel sampling technique, and the remaining part $s_{2,b}$ taken from $U_{2,wd}$. If the sampling is done with a panel technique, the sample parts $s_{2,a}$ and $s_{2,b}$ are the old and new rotation groups, respectively. If the sample is drawn with PRN sampling, $s_{2,a}$ and $s_{2,b}$ consist of units with PRN's that fell in s_1 or did not fall in s_1, respectively. Whether the sample was drawn through PRN sampling or a panel sampling technique, the sample parts can be viewed as two fixed size samples, each drawn with the SI design from their respective subpopulation. We condition on $n_{2,a}$ and $n_{2,b}$ throughout without making it explicit in formulae. With the notation $(k \in s_{2,a})$ we refer to the event that a unit is first included in the first-wave sample(s) from U_1 and then in the second-wave sample taken from what remains of the first-wave sample(s) after dead units have been taken out. The notation $(k \in s_{2,b})$ is analogous. Let $I(k \in s_{2,a}) = 1$ when unit k is included in $s_{2,a}$, otherwise $I(k \in s_{2,a}) = 0$. To derive the overall bias it is convenient to analyse the biases from the sample parts $s_{2,a}$ and $s_{2,b}$. We derive an expression for each of these in section 2.2 and section 2.3, respectively, and in section 2.4 the bias expressions will be amalgamated.

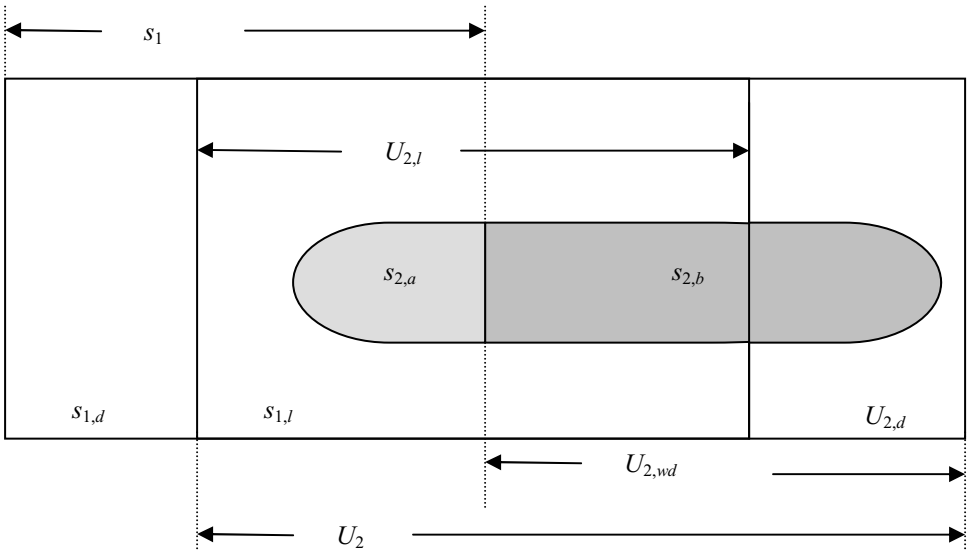


Figure 1. The original survey population, U_1, and its subsets. The grey area represents s_2, the sample from U_2.

2.2 Feed Back Bias from a Sub-sample from the Original Sample

Suppose a sub-sample $s_{2,a}$ is taken from $s_{1,l}$, the live part of the first-wave sample(s). Recall that $y_k = 0$ if k is a dead unit and that $U_2 = U_{2,d} \cup U_{2,l}$. Thus we have $\sum_{s_{2,a}} y_k = \sum_{U_{2,l}} y_k I(k \in s_{2,a}) = \sum_{U_2} y_k I(k \in s_{2,a})$. Assume that $N_{2,l} > 0$. Then we obtain that $\Pr[k \in s_{2,a} \mid n_{1,d}] = n_{2,a} / N_{2,l}$, since a sample of size $n_{2,a}$ is effectively selected from a population of size $N_{2,l}$ with the SI design (through an SI sample from U_1 followed by an SI sample from $U_{2,l}$). Note that a unit k in $s_{2,a}$ must be alive since $U_{2,l}$ consists solely of live units.

Denote the bias of an estimator $\hat{\theta}$ for the parameter θ by $B(\hat{\theta}, \theta)$. Then with respect to the population total $t_y = \sum_{U_2} y_k$, the conditional bias of a general linear estimator $\hat{t}_y^{(s_{2,a})} = \sum_{s_{2,a}} w_k y_k$ based on $s_{2,a}$, with any given w_k's, is

$$\begin{aligned} B\left(\hat{t}_y^{(s_{2,a})}, t_y \middle| n_{1,d}\right) &= \sum_{U_{2,l}} \left\{w_k \Pr\left[k \in s_{2,a} \mid n_{1,d}\right] - 1\right\} y_k \\ &= \sum_{U_{2,l}} \left(\frac{w_k n_{2,a}}{N_{2,l}} - 1\right) y_k \\ &= \sum_{U_2} \left(\frac{w_k n_{2,a}}{N_{2,l}} - 1\right) y_k. \end{aligned} \quad (2)$$

For the sample part $s_{2,a}$, the naive expansion estimator that ignores feed back bias would have weights $w_k = N_2/n_{2,a}$. From (2) the bias of the estimator $\hat{t}_{y\pi}^{(s_{2,a})} = N_2 / n_{2,a} \sum_{s_{2,a}} y_k$ is

$$B\left(\hat{t}_{y\pi}^{(s_{2,a})}, t_y \middle| n_{1,d}\right) = \frac{N_{2,d}}{N_{2,l}} t_y. \quad (3)$$

2.3 Feed Back Bias from a Sample Taken Afresh from the Current Survey Population

Next, we derive the bias arising from the sample part $s_{2,b}$ of size $n_{2,b}$ taken from U_2 through U_{2wd}, see Figure 1. First note that

$$\Pr\left[k \in s_{2,b} \middle| k \in U_{2,\mathrm{wd}}, n_{1,d}\right] = \frac{n_{2,b}}{N_{2,wd}}. \quad (4)$$

From (4) we obtain that the conditional expected value of $\hat{t}_y^{(s_{2,b})} = \sum_{s_{2,b}} w_k y_k$ is

$$\begin{aligned} E\left(\hat{t}_y^{(s_{2,b})} \middle| n_{1,d}\right) &= E\left[\frac{n_{2,b}}{N_{2,wd}} \sum_{U_{2,wd}} w_k y_k \mid n_{1,d}\right] \\ &= \frac{n_{2,b}}{N_{2,wd}} \frac{N_{2,l} - n_{1,l}}{N_{2,l}} \sum_{U_2} w_k y_k. \end{aligned}$$

The second equation above is due to the fact that given $n_{1,d}$, all $N_{2,l}$ live units in U_2 are equally likely to be in $U_{2,wd}$, which has $N_{2,l} - n_{1,l}$ live units. Therefore, the conditional bias of $\hat{t}_y^{(s_{2,b})}$ is

$$B\left(\hat{t}_y^{(s_{2,b})}, t_y \middle| n_{1,d}\right) = \sum_{U_2} \left(\frac{w_k n_{2,b}}{N_{2,wd}} \frac{N_{2,l} - n_{1,l}}{N_{2,l}} - 1\right) y_k. \quad (5)$$

For the expansion estimator $\hat{t}_{y\pi}^{(s_{2,b})}$ with weights $w_k = N_2 / n_{2,b}$ the bias is

$$B\left(\hat{t}_{y\pi}^{(s_{2,b})}, t_y \middle| n_{1,d}\right) = B t_y, \quad (6)$$

where

$$\begin{aligned} B &= \frac{N_2}{N_{2,wd}} \frac{N_{2,l} - n_{1,l}}{N_{2,l}} - 1 \\ &= \frac{N_2\left(N_{2,l} - n_{1,l}\right) - N_{2,l}\left(N_2 - n_{1,l}\right)}{N_{2,wd} N_{2,l}} \\ &= -\frac{N_{2,d} n_{1,l}}{N_{2,l} N_{2,wd}} \\ &= -\frac{N_{2,d}\left(n_1 - n_{1,d}\right)}{N_{2,l}\left(N_1 - n_1\right)}. \end{aligned}$$

The bias is always non-positive since $B \le 0$. It is easy to see that B is an increasing function of $n_{1,d}$ since $N_{2,d} = N_{1,d} - n_{1,d}$, where $N_{1,d}$ is the fixed number of all dead units in U_1. It is also readily seen that the maximum of B is attained when $s_{1,d}$ encompasses all dead units in U_1, that is, when $n_{1,d} = N_{1,d}$ and consequently $N_{2,d} = 0$.

2.4 Feed Back Bias from Sample Parts Combined

Combining (6) with (3) we obtain the overall bias of $\hat{t}_{y\pi} = N_2 / n_2 \sum_{s_2} y_k$ to be

$$\begin{aligned} B\left(\hat{t}_{y\pi}, t_y \middle| n_{1,d}\right) &= E(\hat{t}_{y\pi} \mid n_{1,d}) - t_y \\ &= \frac{N_{2,d}}{N_{2,l}} \left(\frac{n_{2,a}}{n_2} - \frac{n_{2,b}}{n_2} \frac{n_{1,l}}{N_{2,wd}}\right) t_y = \tilde{c} t_y. \end{aligned} \quad (7)$$

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_2 rather than the unknown $N_{2,l}$. The bias from the sample parts $s_{2,a}$ and $s_{2,b}$ will in absolute terms be less than (3) and (6), respectively, if some of the dead units in the samples from U_1 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 (7) with respect to $n_{1,d}$. Thus, unconditionally, we have

$$E\left(\frac{N_2}{n_2}\sum\nolimits_{s_2} y_k\right) - t_y$$
$$= \left[\begin{array}{l} \dfrac{N_{1,d} - E(n_{1,d})}{N_{2,l}}\left(\dfrac{n_{2,a}}{n_2} - \dfrac{n_{2,b}}{n_2}\dfrac{n_1 - E(n_{1,d})}{N_{2,wd}}\right) \\ -\dfrac{n_{2,b}}{n_2 N_{2,l} N_{2,wd}} V(n_{1,d}) \end{array}\right] t_y$$
$$= c t_y \,, \qquad (8)$$

where $E(n_{1,d}) = n_1 N_{1,d} / N_1$ and $V(n_{1,d}) = n_1 N_{1,d} N_{2,l} / N_1^2$.

Lavallée (1996) took an interesting approach to a similar problem with panel survey data. In that paper, the problem of frame update using panels 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 $\Pr[k \in s_{2,a} \mid n_{1,d}]$ and $\Pr[k \in s_{2,b} \mid n_{1,d}]$ separately.

3. Three Simple Strategies and a Simulation Study

3.1 Strategies in the Presence of Feed Back

A strategy, which is referred to as Strategy 1 here, is to feed back, delete the set $s_{1,d}$ from the frame and accept the feed back bias. However, the size of the bias is seldom known. The estimator for Strategy 1 under SI is $\hat{t}_{y\pi} = N_2 / n_2 \sum_{s_2} y_k$ where s_2 is a sample taken from U_2. To obtain Strategy 2, note that if consistent estimates of $N_{2,d}$ and $N_{2,l}$ are available these may be plugged into (7) or (8) and an estimator with favourable properties is obtained:

$$\hat{t}'_{y\pi} = \hat{t}_{y\pi}(1 + \hat{c})^{-1}, \qquad (9)$$

where

$$\hat{c} = (\hat{N}_{2,d} / \hat{N}_{2,l})[n_{2,a} / n_2 - \{n_{2,b}(n_1 - n_{1,d})\} / \{n_2 (N_1 - n_1)\}]$$

for both the conditional and unconditional cases since the term $n_{2,b} V(n_{1,d})(n_2 N_{2,l} N_{2,wd})^{-1}$ in (8) is almost always negligible. The estimates $\hat{N}_{2,d}$ and $\hat{N}_{2,l}$ of the sizes of the domains $U_{2,d}$ and $U_{2,l}$ can be obtained from a sample from the original or current survey population. If more than one sample is drawn, each can provide an unbiased estimate of $N_{2,d}$ (or $N_{2,l}$), all of which can be combined. The minimum variance combined estimator is the sum of the estimators weighted with the reciprocals of their variances. As the following argument shows, we do not expect the bias of (9) to be large:

$$E(\hat{t}'_{y\pi}) = E[\hat{t}_{y\pi}(1 + \hat{c})^{-1}] \approx E(\hat{t}_{y\pi})(1 + c)^{-1}$$
$$= t_y (1 + c)(1 + c)^{-1} = t_y \,.$$

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. 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 $\hat{t}''_{y\pi} = N_1 / n_2 \sum_r y_k$, where r is a sample from the original survey population U_1.

3.2 A Simulation Study

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 1,000 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, page 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, s_1 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 $N_{2,d}$ and $N_{2,l}$ used in Strategy 2 were based on s_1. A second sample, denoted by $s_{2\text{current}}$, 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_1, and we denote it by $s_{2\text{orig}}$. The sample $s_{2\text{current}}$ is pertinent to Strategies 1 and 2 while $s_{2\text{orig}}$ will be used for Strategy 3.

The procedure described in the preceding paragraph was repeated 1,000 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, 1,000 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, $s_{2\text{current}}$, and $s_{2\text{orig}}$). Two values of *start* 2, 0.0 and 0.7, were chosen so as to make the proportion of $s_{2\text{current}}$ that fell in $s_{1,l}$ 100% and 0%, respectively. That is, $n_{2,a}/n_2$ was set to 100% and 0%. Further, to make $n_{2,a}/n_2$ 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 $n_{2,a}/n_2$, 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, $s_{2\text{current}}$, and $s_{2\text{orig}}$;
- how many and which of the dead units that were dead by sample survey sources;
- which of the units that belonged to $s_{1,l}$ and $U_{2,wd}$.

Thus the quantities $s_{1,d}$, $N_{2,d}$ and N_2 vary in the simulations. It seems practical to let them do so rather than controlling them in an experiment with more factors than P_{dead} and $n_{2,a}/n_2$. Hence the results are unconditional, in accordance with (8).

3.3 Results

Table 1 shows the empirical relative bias of Strategies 1 and 2, computed as the straight average of the 1,000 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 1. The empirical bias of Strategy 3 that nevertheless appeared in the simulations reflects the simulation error; it was at most 0.5%. As seen in Table 1, Strategy 2 is virtually unbiased as well. Note that the simulated empirical bias under Strategy 1 is what (8) 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. Figure 2 shows the conditional bias given $n_{1,d}$ for $P_{\text{dead}} = 0.50$ and $n_{2,a}/n_2 = 0\%$. Note that the bias given by (6) is locally well described by the regression line in the figure defined by the OLS fit of the bias conditional on $n_{1,d}$. For example, if $n_{1,d} = 220$, then both $N_{2,d}/N_{2,l}$ and $(n_1 - n_{1,d})/(N_1 - n_1)$ equal 0.56 and $B = -0.31$.

Table 1
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

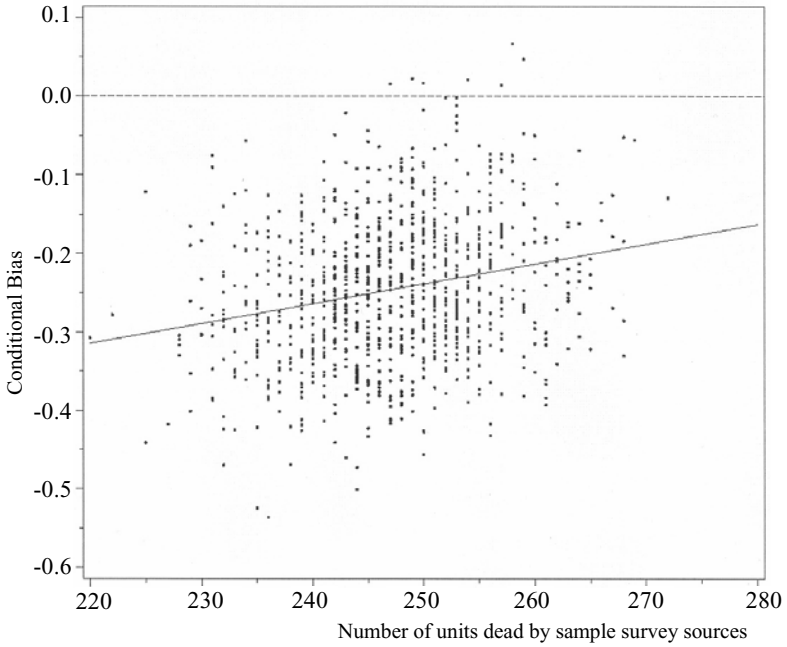


Figure 2. The simulated conditional bias plotted against the number of units dead by sample survey sources, $n_{1,d}$, for $P_{\text{dead}} = 0.50$ and $n_{2,a}/n_2 = 0\%$. An OLS regression line shows the local trend of the conditional bias as a function of $n_{1,d}$.

To assess the bias it helps to look at the coverage probabilities. Table 2 shows the empirical coverage probabilities, based on symmetric 'confidence intervals' with a width of two times the simulated empirical standard deviation of each side of the point estimate. 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.

The variance of the simulated estimates was computed. Tables 3 and 4 show the variance comparisons for Y1 and Y2, respectively, under Strategies 2 and 3 relative to that of Strategy 1. As expected, in all cases Strategy 1 gave a smaller variance than did Strategy 3. Strategy 2 performed well in most cases, but considering the extra complexity of this strategy, the feed back Strategy 1 seems preferable for populations with a small proportion of ineligible units, say 3% or less. However, 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 3 is due to the estimation of $N_{2,d}$ and $N_{d,l}$ combined with the specific details of the simulation.

Table 2
The Coverage Probability in Percentage for Estimating Total of Y1. The First Entry in Each Cell is the Coverage Probability 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 3
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 4
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

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 units 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 through 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'. 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.

Acknowledgements

The authors thank Mark Pont for very useful initial discussions of this topic. They are also most grateful to an associate editor and two referees for very valuable comments. Both authors' research was partially supported by the UK Office for National Statistics and Wang's research was also supported by the U.S. National Cancer Institute (CA 57030). Hedlin was employed by University of Southampton when he took part in this work.

References

Colledge, M.J. (1989). Coverage and classification maintenance issues in economic surveys. In *Panel Surveys,* (Eds., D. Kasprzyk, G.J. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons, Inc., 80-107.

Colledge, M.J. (1995). Frames and business registers: an overview. In *Business Survey Methods,* (Eds., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge and P. Kott). New York: John Wiley & Sons, Inc., 21-47.

Durbin, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling,* (Eds., N.L. Johnson and H. Smith). New York: John Wiley & Sons, Inc., 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, 397-405.

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., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge and P. Kott). New York: John Wiley & Sons, Inc., 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., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge and P. Kott). New York: John Wiley & Sons, Inc., 171-183.

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