Journal of Official Statistics, Vol. 28, No. 1, 2012, pp. 1–27

Capture–Recapture Sampling and Indirect Sampling

Pierre Lavallée[1] *and Louis-Paul Rivest*[2]

Capture–recapture sampling is used to estimate the total number of units in a population of unknown size. It involves two samples selected independently in the target population. The Petersen estimator for the population size depends on the frequencies of the units appearing in the first, the second or both samples. This article considers a generalisation of capture–recapture sampling to cases where the two samples are selected using *indirect sampling* (Lavallée 2002; 2007). The sampling frames for the two samples differ from the target population and the *generalised weight share method* has to be used to determine the sampling weights of units selected in the population through this indirect method. A generalisation of the Petersen estimator to such an indirect sampling scheme is proposed. The sampling properties of this new estimator are investigated. Its application is illustrated through simulations and by discussing two surveys where it could be used.

Key words: Petersen estimator; generalised weight share method; population size; administrative files.

1. Capture–Recapture Sampling

Capture-recapture sampling is used to estimate the total number of units in a population of unknown size. An initial sample s_1 of size n_1 is obtained, and the units in the sample are marked (or identified). A second sample s_2 of size n_2 is obtained *independently*, and the number $n_{1,2}$ of marked units in this sample is recorded. Seber (1982, p. 59) pointed out that if the second sample is a simple random sample from the whole population, the proportion of marked units in this second sample estimates the population proportion. Using this link an estimate of the total number of units in the population is

$$\hat{N}_{Pet} = \frac{n_1 n_2}{n_{1,2}} \tag{1}$$

This estimator is often referred to as the Petersen (or Lincoln-Petersen) estimator. A number of assumptions must be true for expression (1) to be an unbiased estimator of the population size N (see Section 2).

One uses the Petersen estimator in cases where the target population is only partly covered by a set of sampling frames. In practice, this problem occurs with administrative files. If a population is partly covered by two administrative files containing N_1 and N_2

[1] Business Survey Section, Statistics Canada, Ottawa, ON K1A 0T6, Canada.
Email: pierre.lavallee@statcan.gc.ca
[2] Department of Mathematics and Statistics, Université Laval, 1045 rue de la médecine, Québec G1V 0A6, Canada. Email: Louis-Paul.Rivest@mat.ulaval.ca

persons, the Petersen estimator $\hat{N}_{Pet}$ of the population's size is given by the following expression:

$$\hat{N}_{Pet} = \frac{N_1 N_2}{N_{1,2}} \tag{2}$$

where $N_{1,2}$ is the number of members of the population present in both files. Clearly, estimator (2) is valid only if the two files are independent.

It is interesting to note that the problem of estimating the size of a population with incomplete administrative files is related to the problem of *multiple frames*. In this case, the two administrative files constitute two sampling frames – $A1$ and $A2$, say – which are used to measure the target population. This is illustrated by Figure 1 below.

As shown in Figure 1, the target population is only partly covered by frames $A1$ and $A2$, since the union of the members of the two frames does not contain the entire target population. The dots with a triangular shape are not included in either frame; the star dots appear only in frame $A1$; the heart dots appear only in frame $A2$; and the regular round dots are in both frames.

This article examines the case where frames $A1$ and $A2$ cannot be processed in their entirety. Instead, samples are selected from the two administrative files. In other words, the Petersen estimator (2) calculated with the entire populations is regarded as a "census parameter," which we attempt to estimate with two samples, one from each file. With sample s_1 from one file, we obtain an estimate $\hat{N}_1$ of N_1 using the Horvitz-Thompson estimator $\hat{N}_1 = \sum_{k \in s_1} 1/\pi_{1k}$, where π_{1k} is the probability that unit k is selected in sample s_1. Similarly, s_2 is used to produce estimate $\hat{N}_2$. With the units that appear in both samples – which is equivalent to considering the marked units – we estimate N again

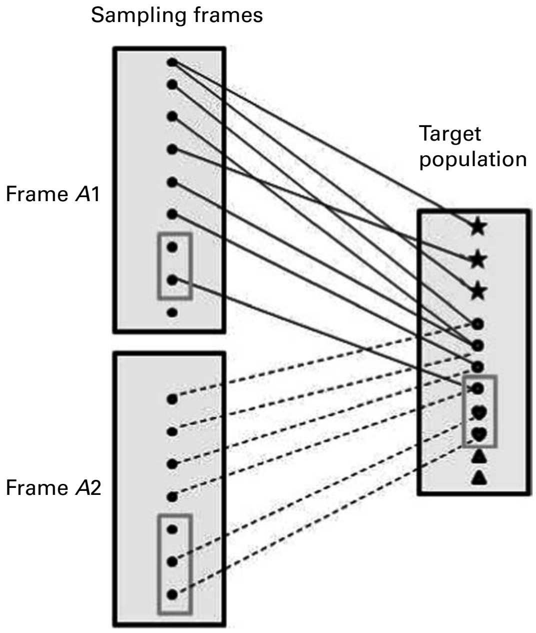

Fig. 1. Multiple frames where samples are represented by grey rectangles

using the Horvitz-Thompson estimator $\hat{N}_{1,2} = \sum_{k \in s_1 \cap s_2} 1/\pi_{1,2,k}$, where $\pi_{1,2,k}$ is the probability that unit k is selected in both samples. In general, the two files are sampled independently, and therefore $\pi_{1,2,k} = \pi_{1k} \times \pi_{2k}$. The final estimator $\hat{\hat{N}}$ of N is obtained by replacing the population characteristics in (2) with their estimators:

$$\hat{\hat{N}} = \frac{\hat{N}_1 \hat{N}_2}{\hat{N}_{1,2}} \tag{3}$$

An alternative to estimator (3) is the classical Petersen estimator (1) calculated with the frequencies of the units sampled from the two files, without accounting for the selection probabilities. When the selection probabilities for the two files are constant, $\pi_{1k} = \pi_1$ and $\pi_{2k} = \pi_2$ for all units, then the estimators are equal. In general, estimator (3) takes unequal selection probabilities into account, thereby correcting for a potential heterogeneity due to unequal capture probabilities in the two frames.

The problem of using administrative files can be generalised to include cases where the files (or frames) do not represent the target population directly, but rather consist of different populations that are related to the target population in some way. In other words, we are attempting to estimate the size of a target population when we do not have a sampling frame for that population, but we have different sampling frames which are related to that population. This is referred to as *indirect sampling* (Lavallée 2002; 2007).

In Section 2 of this article, we take a more detailed look at the Petersen estimator. Section 3 reviews indirect sampling and the generalised weight share method (GWSM). Section 4 introduces capture–recapture sampling in the context of indirect sampling, and describes the generalised capture–recapture estimator. We discuss the estimator's properties in Section 5 and present the results of simulations in Section 6. We follow that with two examples of how the capture–recapture method can be applied to indirect sampling. We will conclude the article by discussing direct application of the GWSM to multiple frames without using the Petersen estimator.

2. Petersen Estimator

The first known use of the Petersen estimator was by Laplace in 1786. To estimate the size of France's population, he multiplied N_1, the number of births for the entire country, by the inverse of the ratio of the number of births to the total population in a number of parishes. That estimation method is named after the Norwegian biologist C.G.J. Petersen, who pioneered the technique to estimate the demographic characteristics of animal populations using marked animals. For more details, see Le Cren (1965) and Otis et al. (1978).

Today, capture–recapture sampling is used in fields such as population biology and epidemiology. A classical example is the estimation of the total number of fish in a lake. The sample s_1 is selected by throwing a net into the lake at random and marking the fish it catches. The next day, s_2 is selected using a second random throw of the net into the lake. Epidemiological applications are concerned with hard to reach or hard to count populations. In that case, s_1 and s_2 are two subsets of the population obtained from different lists. In social statistics, this technique is used to estimate census coverage; the initial sample is the census itself, and the second sample is an independent sample used in a field survey to determine whether each respondent was enumerated in the census (marked) or not. In the context of incomplete sampling frames, such as administrative files, a capture is defined as inclusion in one of the administrative files needed to construct estimator (2).

The model generally used to study the Petersen estimator $\hat{N}_{Pet}$ is the multinomial distribution (see Thompson 2002). In capture–recapture sampling, the population can be divided into four categories: C_{11} is the number of units that appear in both samples s_1 and s_2; C_{10} is the number of units that appear in sample s_1 only; C_{01} is the number of units that appear in sample s_2 only; and C_{00} is the number of units that appear in neither s_1 nor s_2. When the two samples are drawn independently, the probabilities associated with the four categories are $p_{10} = p_1(1 - p_2)$, $p_{01} = (1 - p_1)p_2$, $p_{11} = p_1 p_2$ and $p_{00} = (1 - p_1)(1 - p_2)$, where p_1 and p_2 are the probabilities of being selected in samples 1 and 2 respectively. In other words, we have $(C_{10}, C_{01}, C_{11}, C_{00}) \sim \text{Mult}[N, p_1(1 - p_2), (1 - p_1)p_2, p_1 p_2, (1 - p_1)(1 - p_2)]$ with probability function

$$p(c_{10}, c_{01}, c_{11}, c_{00}) = \frac{N!}{c_{10}!c_{01}!c_{11}!c_{00}!}[p_1(1 - p_2)]^{c_{10}}[(1 - p_1)p_2]^{c_{01}}[p_1 p_2]^{c_{11}} \\ [(1 - p_1)(1 - p_2)]^{c_{00}} \tag{4}$$

provided that $c_{10} + c_{01} + c_{11} + c_{00} = N$. In terms of the quantities N_1, N_2 and $N_{1,2}$ in (2), we have $N_1 = C_{10} + C_{11}$, $N_2 = C_{01} + C_{11}$ and $N_{1,2} = C_{11}$. Note that model (4) applies to the administrative files problem because we do not know *a priori* how many people are in each of the two lists (or files). Under Model (4), we can show that estimator (2) is asymptotically unbiased; that is, $E_\xi(\hat{N}_{Pet}) \approx N$, where the subscript ξ indicates that the expected value is calculated under (4). For more details, see Chapman (1951). The estimator's asymptotic variance is given by

$$V_\xi(\hat{N}_{Pet}) = N\frac{(1 - p_1)(1 - p_2)}{p_1 p_2} \tag{5}$$

see Sekar and Deming (1949) and Thompson (2002). In the case of capture–recapture sampling where the sizes n_1 and n_2 of s_1 and s_2 are fixed, the frequencies C_{ij} have a hypergeometric distribution (see Seber 1982).

Seber (1982, p. 59) presents a set of conditions for estimator (1) to be unbiased. One condition is that a unit selected in the first sample and a unit that is not selected must have the same probability of being selected in the second sample. For example, a reaction to the capture event that increases (or decreases) the probability of being recaptured causes (1) to be biased, which makes it necessary to use an estimator that takes a behavioural effect into account (see Seber 1982, p. 318). The estimator $\hat{N}_{Pet}$ is sensitive to a heterogeneity that gives some units a greater chance than others of being selected in both capture events (see Seber 1970). In the case of the administrative files used to construct (2), people from modest socio-economic backgrounds, for example, may have lower coverage rates. Some geographic areas may also be less well represented in the files than other areas. This heterogeneity can be corrected for by splitting the population into groups that are relatively homogeneous with respect to the heterogeneity variable, and applying the Petersen estimator to each group separately. For epidemiological and biological examples of application of this approach, see Hook and Regal (1993) and Rivest et al. (1995). Alho

(1990) and Chen and Lloyd (2000) suggest methods to include the heterogeneity variable in the model. The stratification described by Plante et al. (1998) provides an alternative way of correcting for heterogeneity. Note that capture heterogeneity associated with unobserved variables can also be dealt with by using capture-recapture estimators based on three or more files.

3. Indirect Sampling and the Generalised Weight Share Method (GWSM)

Indirect sampling consists in selecting a sample from a frame U^A for the purpose of surveying a target population U^B that is not the population represented by the frame, but is related to that population. More formally, suppose a sample s^A of n^A units is selected from a population U^A of N^A units using a particular sample design. Let π_j^A be the selection probability of unit j. We assume that $\pi_j^A > 0$ for all $j \in U^A$. We also assume that the target population U^B contains N^B units. We are interested in estimating the total $Y^B = \sum_{k=1}^{N^B} y_k$ in population U^B for the variable of interest y.

We assume that there is a *link* (or *relationship*) between the units j of population U^A and the units k of population U^B. That link is identified by indicator variable $l_{j,k}$, where $l_{j,k} = 1$ if there is a link between unit $j \in U^A$ and unit $k \in U^B$, and 0 if not. Note that there may be cases where there is no link between a unit j of population U^A and the units k of target population U^B, that is, $L_j^A = \sum_{k=1}^{N^B} l_{j,k} = 0$.

For each unit j selected in s^A, we identify the units k of U^B that have a nonzero link with j, that is, $l_{j,k} = 1$. If $L_j^A = 0$ for a unit j of s^A, there is simply no unit of U^B identified with that unit j; this affects the efficiency of sample s^A but does not cause bias. For each unit k identified, we measure a particular variable of interest y_k and the number of links L_k^B between unit k of U^B and population U^A. Let s^B be the set of n^B units of U^B identified by units $j \in s^A$.

For target population U^B, we want to estimate the total Y^B. Estimating that total is a major challenge if the links between the units of the two populations are not one-to-one. The problem is due primarily to the difficulty of associating a selection probability, or an estimation weight, with the units of the target population that is surveyed. The GWSM, as described in Lavallée (1995; 2002; 2007), assigns an estimation weight w_k to each surveyed unit k. The method relies on sample s^A and the links between U^A and U^B to estimate the total Y^B. To estimate the total Y^B for target population U^B, we can use the estimator

$$\hat{Y}^B = \sum_{k \in s^B} w_k y_k \tag{6}$$

The GWSM is an extension of the *weight share method* described by Ernst (1989) in the context of longitudinal household surveys. It can be regarded as a generalisation of *network sampling* and *adaptive cluster sampling*, see Thompson (2002) and Thompson and Seber (1996).

In formal terms, the GWSM assigns a weight w_k to each unit k in s^B

$$w_k = \frac{1}{L_k^B} \sum_{j=1}^{N^A} l_{j,k} \frac{t_j}{\pi_j^A} \tag{7}$$

where $t_j = 1$ if $j \in s^A$ and 0 if not and $L_k^B = \sum_{j=1}^{N^A} l_{j,k}$. It is important to note that if unit j of U^A is not selected, we do not need to know its selection probability π_j^A, a key point in the GWSM. In addition, for the GWSM to be unbiased, we must have $L_k^B > 0$; in other words, each unit k of U^B must have at least one link with U^A. As shown in Lavallée (1995), $\hat{Y}^B$ can also be written as

$$\hat{Y}^B = \sum_{j=1}^{N^A} \frac{t_j}{\pi_j^A} \sum_{k=1}^{N^B} l_{j,k} \frac{y_k}{L_k^B} = \sum_{j=1}^{N^A} \frac{t_j}{\pi_j^A} Z_j \tag{8}$$

Using the latter expression, we can easily show that the GWSM is design-unbiased. The variance of $\hat{Y}^B$ is computed directly with

$$V_p(\hat{Y}^B) = \sum_{j=1}^{N^A} \sum_{j'=1}^{N^A} \frac{\left(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A\right)}{\pi_j^A \pi_{j'}^A} Z_j Z_{j'} \tag{9}$$

where $\pi_{jj'}^A$ is the joint selection probability of units j and j', and $\pi_{jj}^A = \pi_j^A$. Methods for calculating the $\pi_{jj'}^A$ under various sample designs are given in Särndal, Swensson, and Wretman (1992). Subscript p indicates that the variance is being calculated relative to the sample design.

An unbiased estimate of the variance $V_p(\hat{Y}^B)$ is given by

$$\hat{V}_p(\hat{Y}^B) = \sum_{j=1}^{N^A} \sum_{j'=1}^{N^A} \frac{\left(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A\right)}{\pi_{jj'}^A \pi_j^A \pi_{j'}^A} t_j Z_j t_{j'} Z_{j'} \tag{10}$$

Another estimator of the variance $V_p(\hat{Y}^B)$ can be developed in the form proposed by Yates and Grundy (1953).

It should be noted that the indirect sampling operation considered in this section assumes complete response. That is, no nonresponse occurres during the collection process. Although this is unrealistic in practice, adjusting for nonresponse is out of the scope of this article. We can however mention that with indirect sampling, there are three types of nonresponse: (i) nonresponse within s^A; (ii) nonresponse within s^B; (iii) errors in the identification of the links $l_{j,k}$. Nonresponse can be treated by adjusting the weights of estimator (6) (or (8)) according to each of the three types of nonresponse. For more details, see Lavallée (2002; 2007).

4. Generalised Capture–Recapture Estimator

An unbiased (or approximately unbiased) estimator of the size N^B of the target population U^B can be developed using indirect sampling based on estimator (2). The two administrative files $A1$ and $A2$ constitute populations U^{A1} and U^{A2}, whose sizes are N^{A1} and N^{A2}, respectively. The target population U^B is different from the sampling frames U^{A1} and U^{A2}. In this general context, we suppose that, using particular sample designs, samples s^{A1} of n^{A1} units and s^{A2} of n^{A2} units are selected from populations U^{A1} and U^{A2}, respectively. Let s^{B1} (s^{B2}) be the units in U^B with at least one link to a unit in s^{A1} (s^{A2}). It is useful to differentiate $l_{j,k}^{A1}$, which takes a value of 1 when unit j of U^{A1} has a link with unit k of U^B and 0 otherwise, from $l_{j,k}^{A2}$, which describes a link between U^{A2} and U^B. We have $\sum_{j=1}^{N^{A1}} l_{j,k}^{A1} = L_k^{A1}$, $\sum_{j=1}^{N^{A2}} l_{j,k}^{A2} = L_k^{A2}$, where L_k^{A1} and L_k^{A2} represent the total number of links that unit k of U^B has with U^{A1} and U^{A2}, respectively. Then three indirect sampling operations are conducted: one from U^{A1} to U^B; another from U^{A2} to U^B; and a third from U^{A1} and U^{A2} to U^B. This last one involves only the units of $s^{B1} \cap s^{B2}$ *that have at least one link with both samples* s^{A1} and s^{A2}.

Let $U_{A1}^B = \{k \in U^B | \exists j \in U^{A1}, l_{j,k}^{A1} \neq 0\}$ be the subpopulation of U^B whose units have at least one link with a unit in U^{A1}; the size of this population is N_{A1}^B. Let $U_{A2}^B = \{k \in U^B | \exists j \in U^{A2}, l_{j,k}^{A2} \neq 0\}$ of size N_{A2}^B. In addition, let $U_{A1,A2}^B = U_{A1}^B \cap U_{A2}^B$ be the subpopulation of U^B whose units have a link with both U^{A1} and U^{A2}; the size of $U_{A1,A2}^B$ is $N_{A1,A2}^B$. If we could conduct a census of U^{A1} and U^{A2}, we could estimate N^B with $N_{A1}^B \times N_{A2}^B / N_{A1,A2}^B$.

Let π_j^{A1} and π_j^{A2} be the selection probabilities of unit j of U^{A1} and U^{A2}. To develop an estimator for $N_{A1}^B \times N_{A2}^B / N_{A1,A2}^B$, we first obtain estimators $\hat{N}_{A1}^B$ and $\hat{N}_{A2}^B$ from samples s^{A1} and s^{A2} using (6) (or (8)). Second, we estimate $N_{A1,A2}^B$ using only those units of $s^{B1} \cap s^{B2}$. To do so, we need to alter the GWSM presented in Section 3. Each unit k identified by both samples s^{A1} and s^{A2} is assigned the following weight $w_k^{A1,A2}$

$$w_k^{A1,A2} = w_k^{A1} w_k^{A2} = \left(\frac{1}{L_k^{A1}} \sum_{j=1}^{N^{A1}} l_{j,k}^{A1} \frac{t_j^{A1}}{\pi_j^{A1}} \right) \cdot \left(\frac{1}{L_k^{A2}} \sum_{j=1}^{N^{A2}} l_{j,k}^{A2} \frac{t_j^{A2}}{\pi_j^{A2}} \right) \tag{11}$$

Finally $w_k^{A1,A2} = 0$ unless $l_{j,k}^{A1} = 1$ and $l_{j',k}^{A2} = 1$ for some units j and j' of s^{A1} and s^{A2}, respectively. Only the $N_{A1,A2}^B$ units of $U_{A1,A2}^B$ can be assigned a positive weight $w_k^{A1,A2}$ in this estimator. We therefore have the estimator

$$\hat{N}_{A1,A2}^B = \sum_{k=1}^{N_{A1,A2}^B} \left(\frac{1}{L_k^{A1}} \sum_{j=1}^{N^{A1}} l_{j,k}^{A1} \frac{t_j^{A1}}{\pi_j^{A1}} \right) \cdot \left(\frac{1}{L_k^{A2}} \sum_{j=1}^{N^{A2}} l_{j,k}^{A2} \frac{t_j^{A2}}{\pi_j^{A2}} \right) \tag{12}$$

Lastly, on the basis of (2), we construct the *generalised capture–recapture estimator* (in French, *estimateur par capture–recapture généralisé*, CReG) of N^B

$$\hat{N}_{CReG}^B = \frac{\hat{N}_{A1}^B \hat{N}_{A2}^B}{\hat{N}_{A1,A2}^B} \tag{13}$$

It is important to note that an estimator similar to (13) can be developed in an even more general context: the estimation of the total Y^B for the variable of interest y. We may want to estimate not only a population size such as N^B, but also a total Y^B.

As in the process used to construct (6), to estimate the total Y_{A1}^B for the units of U^B that have a link with U^{A1} based on s^{A1}, we use

$$\hat{Y}_{A1}^B = \sum_{k \in s^{B1}} w_k^{A1} y_k = \sum_{k=1}^{N^B} y_k \frac{1}{L_k^{A1}} \sum_{j=1}^{N^{A1}} l_{j,k}^{A1} \frac{t_j^{A1}}{\pi_j^{A1}} \tag{14}$$

where w_k^{A1} is the GWSM weight for unit k given by (7). The estimator of Y_{A2}^B, constructed with s^{B2}, has the same form.

We now turn to $Y^B_{A1,A2}$, the sum of y for the $N^B_{A1,A2}$ units of U^B that have a link with both U^{A1} and U^{A2}. We estimate that total using only those units of U^B that have a link with both samples s^{A1} and s^{A2}

$$\hat{Y}^B_{A1,A2} = \sum_{k=1}^{N^B_{A1,A2}} \left(\frac{1}{L^{A1}_k} \sum_{j=1}^{N^{A1}} l^{A1}_{j,k} \frac{t^{A1}_j}{\pi^{A1}_j} \right) \cdot \left(\frac{1}{L^{A2}_k} \sum_{j=1}^{N^{A2}} l^{A2}_{j,k} \frac{t^{A2}_j}{\pi^{A2}_j} \right) y_k \tag{15}$$

Lastly, we construct the CReG estimator of the total Y^B in U^B

$$\hat{Y}^B_{CReG} = \frac{\hat{Y}^B_{A1} \hat{Y}^B_{A2}}{\hat{Y}^B_{A1,A2}} \tag{16}$$

Note that as in the case of estimator (3), if the selection probabilities are equal within each frame U^{A1} and U^{A2} (that is, $\pi^{A1}_j = f^{A1}$ and $\pi^{A2}_j = f^{A2}$), we can cancel out π^{A1}_j and π^{A2}_j in the numerator and denominator of (16).

In the presence of nonresponse, the CReG estimator can be adjusted for correcting: (i) nonresponse within s^{A1} and s^{A2}; (ii) nonresponse within s^{B1} and s^{B2}; (iii) errors in the identification of the links $l_{j,k}$ in both indirect sampling operations. In particular, estimators (12) and (15) can be adjusted as the ones for the GWSM (see Lavallée 2002; 2007).

5. Properties of the CReG Estimator

5.1. Asymptotic Framework

In capture–recapture experiments, the limit properties of the estimators – convergence in probability and convergence in distribution – are proved by having the population size N tending to infinity. We say that $\hat{N}$ is convergent if $\hat{N}/N$ tends to 1 in probability as $N \longrightarrow \infty$. It can be shown that the limit distribution of $\sqrt{N}(\hat{N}/N - 1)$ when N tends to infinity is normal. Its variance determines the asymptotic variance of $\hat{N}$.

To rigorously determine the properties of the estimator studied in this article, we need to consider a sequence of populations $\left\{ U^B_N : N \geq N_0 \right\}$ in which the size of U^B_N is N^B. A value of y, the variable of interest, is associated with each unit of U^B_N. The creation of the two frames from which U^B_N is indirectly sampled must satisfy the assumptions below. Those assumptions define the model ξ mentioned in Section 2.

Under model ξ, the frames used to sample U^B_N indirectly are formed using the following pseudo-sampling procedure:

1. From U^B_N, we select S^B_{1N}, a Bernoulli sample in which a unit's selection probability is p_1. Let the size of the sample be N^B_{A1}.
2. With each element k of S^B_{1N}, we associate a set of $L^B_{1,k}$ links, where $L^B_{1,k} > 0$, using a particular process. (This process does not have to be formally defined since all the properties (biases and variances) are calculated by conditioning on the links.) The population U^{A1}_N is formed from the union, for all elements of S^B_{1N}, of those sets of links. Note that there may be units j of U^{A1}_N with $l^{A1}_{j,k} = 1$ for more than one unit k of U^B_N, and therefore the size N_{A1} of U^{A1}_N satisfies $N_{A1} \leq \sum_{k=1}^{N^B_{A1}} L^B_{1,k}$.

3. To form U_N^{A2}, we follow the same procedure with a Bernoulli sample S_{2N}^B of size N_{A2}^B and selection probability p_2.
4. The selection probabilities in forming U_N^{A1} and U_N^{A2} are independent of y.

This procedure constructs sequences of populations $\{U_N^{A1}\}$ and $\{U_N^{A2}\}$ that can be used to sample the elements of $\{U_N^B\}$ indirectly. The number of elements of U_N^B that belong to both S_{1N}^B and S_{2N}^B is $N_{A1,A2}^B$. The joint distribution under model ξ of population sizes N_{A1}^B, N_{A2}^B and $N_{A1,A2}^B$ satisfies

$$\left[\left(N_{A1}^B - N_{A1,A2}^B\right), \left(N_{A2}^B - N_{A1,A2}^B\right), N_{A1,A2}^B\right] \sim \text{Mult}\left(N^B, p_1(1-p_2), p_2(1-p_1), p_1p_2\right)$$

This is the same as model (4) in Section 2. Under model ξ, we have the following convergences in probability as N tends to infinity:

$$\frac{N_{A1,A2}^B}{N^B} \xrightarrow{\text{Pr}} p_1p_2, \ \frac{N_{A1}^B}{N^B} \xrightarrow{\text{Pr}} p_1 \text{ and } \frac{N_{A2}^B}{N^B} \xrightarrow{\text{Pr}} p_2 \tag{17}$$

In the general framework, a value of the variable of interest y is associated with each unit of U_N^B. It is assumed that y is independent of the random variables associated with the formation of the administrative files for indirect sampling.

5.2. Consistency

Let μ^B be the limit of $\overline{Y}_N^B$, the mean of y for U_N^B when N tends to infinity. Because U_N^{A1} and U_N^{A2} are constructed by independent binomial sampling, the means $\overline{Y}_{1,N}^B$, $\overline{Y}_{2,N}^B$ and $\overline{Y}_{1,2,N}^B$ of y for the sets S_{1N}^B, S_{2N}^B and $S_{1,2,N}^B = S_{1N}^B \cap S_{2N}^B$ all converge to μ^B when N tends to infinity.

Consider

$$\hat{Y}_{CReG}^B = \frac{\hat{Y}_{A1}^B \hat{Y}_{A2}^B}{\hat{Y}_{A1,A2}^B} = \frac{N_{A1}^B N_{A2}^B}{N_{A1,A2}^B} \times \frac{\hat{\overline{Y}}_{A1}^B \hat{\overline{Y}}_{A2}^B}{\hat{\overline{Y}}_{A1,A2}^B} \tag{18}$$

where $\hat{\overline{Y}}_{A1}^B = \hat{Y}_{A1}^B / N_{A1}^B$, $\hat{\overline{Y}}_{A2}^B = \hat{Y}_{A2}^B / N_{A2}^B$ and $\hat{\overline{Y}}_{A1}^B = \hat{Y}_{A1,A2}^B / N_{A1,A2}^B$. Expression (18) shows that $\hat{Y}_{CReG}^B / N^B$ converges to μ^B when N tends to infinity. The means $\hat{\overline{Y}}_{A1}^B$, $\hat{\overline{Y}}_{A2}^B$ and $\hat{\overline{Y}}_{A1,A2}^B$ all converge to μ^B because the GWSM yields unbiased estimates of the underlying totals, Y_{A1}^B, Y_{A2}^B and $Y_{A1,A2}^B$ (which are in fact equal to $Y_{1,N}^B$, $Y_{2,N}^B$ and $Y_{1,2,N}^B$, for a given N). Moreover, by virtue of (17),

$$\frac{N_{A1}^B N_{A2}^B}{N_{A1,A2}^B N^B} \xrightarrow{\text{Pr}} 1$$

Hence, estimator $\hat{Y}_{CReG}^B / N^B$ is consistent for the estimation of $\overline{Y}^B = Y^B / N^B$.

5.3. Bias and Variance

In Appendix A, we show that estimator (15) is design-unbiased, i.e., $E_p\left(\hat{Y}_{A1,A2}^B\right) = Y_{A1,A2}^B$. We can also show that estimator (16) is asymptotically unbiased, with respect to model ξ and the design, for the estimation of Y^B; in other words, $E_\xi E_p\left(\hat{Y}_{CReG}^B\right) - Y^B \approx 0$.

With regard to calculating the variance of estimator (16), we can see that the proposed model is similar to two-phase sampling. Phase 1 is the pseudo-sampling under model ξ, and phase 2 is the indirect sampling of units associated with the two administrative files "created" in phase 1. We begin with the identity

$$V\left(\hat{Y}^B_{CReG}\right) = E_\xi V_p\left(\hat{Y}^B_{CReG}\right) + V_\xi E_p\left(\hat{Y}^B_{CReG}\right) \tag{19}$$

We will write $\delta_p\hat{Y} = (\hat{Y} - Y)/Y$. Using this notation, estimator (16) can be written as

$$\hat{Y}^B_{CReG} \approx \frac{Y^B_{A1}Y^B_{A2}}{Y^B_{A1,A2}}\Bigg[1 + \delta_p\hat{Y}^B_{A1} + \delta_p\hat{Y}^B_{A2} + \delta_p\hat{Y}^B_{A1}\delta_p\hat{Y}^B_{A2} - \delta_p\hat{Y}^B_{A1,A2} \\ - \delta_p\hat{Y}^B_{A1}\delta_p\hat{Y}^B_{A1,A2} - \delta_p\hat{Y}^B_{A2}\delta_p\hat{Y}^B_{A1,A2} + \delta_p^2\hat{Y}^B_{A1,A2}\Bigg] \tag{20}$$

From (20) we get

$$V_p\left(\hat{Y}^B_{CReG}\right) \approx \left(\frac{Y^B_{A1}Y^B_{A2}}{Y^B_{A1,A2}}\right)^2\left(\frac{V_p\left(\hat{Y}^B_{A1}\right)}{\left(\hat{Y}^B_{A1}\right)^2} + \frac{V_p\left(\hat{Y}^B_{A2}\right)}{\left(Y^B_{A2}\right)^2} + \frac{V_p\left(\hat{Y}^B_{A1,A2}\right)}{\left(Y^B_{A1,A2}\right)^2} \\ -2\frac{Cov_p\left(\hat{Y}^B_{A1}, \hat{Y}^B_{A1,A2}\right)}{Y^B_{A1}Y^B_{A1,A2}} - 2\frac{Cov_p\left(\hat{Y}^B_{A2}, \hat{Y}^B_{A1,A2}\right)}{Y^B_{A2}Y^B_{A1,A2}}\right) \tag{21}$$

Taking the expected value of (20), we have

$$E_p\left(\hat{Y}^B_{CReG}\right) \approx \frac{Y^B_{A1}Y^B_{A2}}{Y^B_{A1,A2}} \tag{22}$$

Substituting (21) and (22) in (19), we obtain

$$V\left(\hat{Y}^B_{CReG}\right) \approx E_\xi\left[\left(\frac{Y^B_{A1}Y^B_{A2}}{Y^B_{A1,A2}}\right)^2\left(\frac{V_p\left(\hat{Y}^B_{A1}\right)}{\left(Y^B_{A1}\right)^2} + \frac{V_p\left(\hat{Y}^B_{A2}\right)}{\left(Y^B_{A2}\right)^2} + \frac{V_p\left(\hat{Y}^B_{A1,A2}\right)}{\left(Y^B_{A1,A2}\right)^2} \\ -2\frac{Cov_p\left(\hat{Y}^B_{A1}, \hat{Y}^B_{A1,A2}\right)}{Y^B_{A1}Y^B_{A1,A2}} - 2\frac{Cov_p\left(\hat{Y}^B_{A2}, \hat{Y}^B_{A1,A2}\right)}{Y^B_{A2}Y^B_{A1,A2}}\right)\right] + V_\xi\left(\frac{Y^B_{A1}Y^B_{A2}}{Y^B_{A1,A2}}\right) \tag{23}$$

The variances $V_p\left(\hat{Y}^B_{A1}\right)$ and $V_p\left(\hat{Y}^B_{A2}\right)$ are given by (9), with the corresponding notation. For the variance $V_p\left(\hat{Y}^B_{A1,A2}\right)$ and the covariances $Cov_p\left(\hat{Y}^B_{A1}, \hat{Y}^B_{A1,A2}\right)$ and $Cov_p\left(\hat{Y}^B_{A2}, \hat{Y}^B_{A1,A2}\right)$, see Appendix B.

We have $\delta_\xi Y^B_{A1} = \left(Y^B_{A1} - p_1N^B\mu^B\right)/\left(p_1N^B\mu^B\right)$ and similar expressions for $\delta_\xi Y^B_{A2}$ and $\delta_\xi Y^B_{A1A2}$ where $\mu^B = E_\xi\left(\overline{Y}^B_N\right)$ is the limit of $\overline{Y}^B_{1,N}$, $\overline{Y}^B_{2,N}$ and $\overline{Y}^B_{1,2,N}$. A derivation similar to that of (21) allows us to approximate the second term of (23) as follows

$$V_\xi\left(\frac{Y^B_{A1}Y^B_{A2}}{Y^B_{A1,A2}}\right) \approx (N^B\mu)^2\left(\frac{V_\xi\big(Y^B_{A1}\big)}{(p_1N^B\mu)^2}+\frac{V_\xi\big(Y^B_{A2}\big)}{(p_2N^B\mu)^2}+\frac{V_\xi\Big(Y^B_{A1,A2}\Big)}{(p_1p_2N^B\mu)^2}\right.$$
$$\left.-2\frac{Cov_\xi\Big(Y^B_{A1},Y^B_{A1,A2}\Big)}{p_2(p_1N^B\mu)^2}-2\frac{Cov_\xi\Big(Y^B_{A2},Y^B_{A1,A2}\Big)}{p_1(p_2N^B\mu)^2}\right) \tag{24}$$

We can rewrite (24) as follows:

$$V_\xi\left(\frac{Y^B_{A1}Y^B_{A2}}{Y^B_{A1,A2}}\right) \approx \left(\frac{V_\xi\big(Y^B_{A1}\big)}{p_1^2}+\frac{V_\xi\big(Y^B_{A2}\big)}{p_2^2}+\frac{V_\xi\Big(Y^B_{A1,A2}\Big)}{p_1^2p_2^2}\right.$$
$$\left.-2\frac{Cov_\xi\Big(Y^B_{A1},Y^B_{A1,A2}\Big)}{p_1^2p_2}-2\frac{Cov_\xi\Big(Y^B_{A2},Y^B_{A1,A2}\Big)}{p_1p_2^2}\right) \tag{25}$$

Now, under model ξ in Section 5.1, the frames U_N^{A1} and U_N^{A2} used to sample U_N^B indirectly are formed with Bernoulli samples in which a unit's selection probability is p_1 and p_2, respectively. Hence, we can write $Y^B_{A1}/p_1 = \sum_{k\in U_N^B} u_k^{A1}y_k/p_1$, where $u_k^{A1}=1$ if $k\in S^B_{1,N}$, 0 if not. Since it is a Bernoulli sample, the variance $V_\xi\big(Y^B_{A1}/p_1\big)$ is simply

$$V_\xi\big(Y^B_{A1}/p_1\big)=\sum_{k\in U_N^B}\frac{(1-p_1)}{p_1}y_k^2 \tag{26}$$

(see Särndal, Swensson, and Wretman 1992). We proceed in the same way for $V_\xi\big(Y^B_{A2}\big)=\sum_{k\in U_N^B}(1-p_2)y_k^2/p_2$ and $V_\xi\Big(Y^B_{A1,A2}\Big)=\sum_{k\in U_N^B}(1-p_1p_2)y_k^2/(p_1p_2)$. In addition, we can show that $Cov_\xi\Big(Y^B_{A1},Y^B_{A1,A2}\Big)/\big(p_1^2p_2\big)=Var_\xi\big(Y^B_{A1}/p_1\big)$ and $Cov_\xi\Big(Y^B_{A2},Y^B_{A1,A2}\Big)/\big(p_1p_2^2\big)=Var_\xi\big(Y^B_{A2}/p_2\big)$. Combining these results, we get

$$V_\xi\left(\frac{Y^B_{A1}Y^B_{A2}}{Y^B_{A1,A2}}\right)\approx\sum_{k\in U_N^B}y_k^2\frac{(1-p_1)(1-p_2)}{p_1p_2} \tag{27}$$

Note that if $y_k=1$ for all units k of U_N^B, the variance (27) is the variance of the Petersen estimator given by (5).

Lastly, the variance of $\hat{Y}^B_{CReG}$ is asymptotically given by

$$
\begin{aligned}
V(\hat{Y}^B_{CReG}) \approx & E_\xi\left[\left(\frac{Y^B_{A1}Y^B_{A2}}{Y^B_{A1,A2}}\right)^2\left(\frac{V_p\left(\hat{Y}^B_{A1}\right)}{\left(Y^B_{A1}\right)^2}+\frac{V_p\left(\hat{Y}^B_{A2}\right)}{\left(Y^B_{A2}\right)^2}+\frac{V_p\left(\hat{Y}^B_{A1,A2}\right)}{\left(Y^B_{A1,A2}\right)^2}\right.\right. \\
& \left.\left.-2\frac{Cov_p\left(\hat{Y}^B_{A1},\hat{Y}^B_{A1,A2}\right)}{Y^B_{A1}Y^B_{A1,A2}}-2\frac{Cov_p\left(\hat{Y}^B_{A2},\hat{Y}^B_{A1,A2}\right)}{Y^B_{A2}Y^B_{A1,A2}}\right)\right] \\
& +\sum_{k\in U^B_N} y_k^2\frac{(1-p_1)(1-p_2)}{p_1p_2}
\end{aligned}
\tag{28}
$$

Let $\phi_k = y_k^2(1-p_1)(1-p_2)/(p_1p_2)$. Since the variance $V_\xi\left(Y^B_{A1}Y^B_{A2}/Y^B_{A1,A2}\right) \approx \sum_{k\in U^B_N}\phi_k = \Phi^B_N$ and since it represents only an unknown total within the target population U^B_N, we can estimate that variance using the CReG estimator given by (16). Thus we have

$$
\hat{V}_\xi\left(\hat{Y}^B_{CReG}\right)=\frac{\hat{\Phi}^B_{A1}\hat{\Phi}^B_{A2}}{\hat{\Phi}^B_{A1,A2}}
\tag{29}
$$

where

$$
\hat{\Phi}^B_{A1}=\sum_{k\in s^{B1}} w_k^{A1}\hat{\phi}_k
\tag{30}
$$

$$
\hat{\Phi}^B_{A2}=\sum_{k\in s^{B2}} w_k^{A2}\hat{\phi}_k
\tag{31}
$$

$$
\hat{\Phi}^B_{A1,A2}=\sum_{k\in s^{B1}\cap s^{B2}}\left(\frac{1}{L_k^{A1}}\sum_{j=1}^{N^{A1}} l^1_{j,k}\,\frac{t_j^{A1}}{\pi_j^{A1}}\right)\cdot\left(\frac{1}{L_k^{A2}}\sum_{j=1}^{N^{A2}} l^2_{j,k}\,\frac{t_j^{A2}}{\pi_j^{A2}}\right)\hat{\phi}_k
\tag{32}
$$

and $\hat{\phi}_k$ is a plug-in estimator for ϕ_k. On the basis of (28) and (29), an estimator of the variance of $\hat{Y}^B_{CReG}$ is given by

$$
\begin{aligned}
\hat{V}\left(\hat{Y}^B_{CReG}\right) \approx & \left(\frac{\hat{Y}^B_{A1}\hat{Y}^B_{A2}}{\hat{Y}^B_{A1,A2}}\right)^2\left(\frac{\hat{V}_p\left(\hat{Y}^B_{A1}\right)}{\left(\hat{Y}^B_{A1}\right)^2}+\frac{\hat{V}_p\left(\hat{Y}^B_{A2}\right)}{\left(\hat{Y}^B_{A2}\right)^2}+\frac{\hat{V}_p\left(\hat{Y}^B_{A1,A2}\right)}{\left(\hat{Y}^B_{A1,A2}\right)^2}\right. \\
& \left.-2\frac{\hat{C}ov_p\left(\hat{Y}^B_{A1},\hat{Y}^B_{A1,A2}\right)}{\hat{Y}^B_{A1}\hat{Y}^B_{A1,A2}}-2\frac{\hat{C}ov_p\left(\hat{Y}^B_{A2},\hat{Y}^B_{A1,A2}\right)}{\hat{Y}^B_{A2}\hat{Y}^B_{A1,A2}}\right)+\hat{V}_\xi\left(\hat{Y}^B_{CReG}\right)
\end{aligned}
\tag{33}
$$

6. Simulations

In this section, we present a small simulation study of the empirical properties of estimators (13) and (16). We consider the case in which the units in files $A1$ and $A2$ have a link with no more than one unit in file B. We therefore assume that $L_j^{A1} = \sum_{k=1}^{N^B} l_{j,k}^{A1}$ and $L_j^{A2} = \sum_{k=1}^{N^B} l_{j,k}^{A2}$ are equal to 0 or 1 for all $j \in U^{A1}$ and all $j \in U^{A2}$. We also consider the estimators obtained without using the sample designs for U^{A1} and U^{A2}. They are the Petersen estimator for N^B and the generalised Petersen estimator for Y^B

$$\hat{Y}_{Pet}^{B} = \frac{\sum\limits_{k \in s^{B1}} y_k \sum\limits_{k \in s^{B2}} y_k}{\sum\limits_{k \in s^{B1} \cap s^{B2}} y_k} \tag{34}$$

In the simulations, L_k^{A1} and L_k^{A2}, the numbers of links that unit k of U^B has with U^{A1} and U^{A2} range between 0 and 4. The joint distribution of $\left(L_k^{A1}, L_k^{A2}\right)$ is determined by a 5×5 matrix that gives the probabilities of the $\left(L_k^{A1}, L_k^{A2}\right)$ pairs. Tables 1 and 2 show two matrices used in the simulations. For both matrices, the selection probabilities for Bernoulli sampling of model ξ are $p_1 = p_2 = 0.8$. They satisfy the assumption of independence between the two samples taken under model ξ; the 2×2 tables generated by lumping together the 1, 2, 3, and 4 values of L_k^{A1} and L_k^{A2} satisfy the assumption of independence between the rows and columns. The matrix $\mathbf{M}_1$ yields a correlation of about 0.2 between L_k^{A1} and L_k^{A2}, and for $\mathbf{M}_2$, the correlation is 0.75. After L_k^{A1} and L_k^{A2} were simulated, the value of y_k was generated by a gamma distribution with a shape parameter of 10 and a scale parameter proportional to $\left(L_k^{A1} + 1\right) \times \left(L_k^{A2} + 1\right)$.

In addition to the matrices in Tables 1 and 2, matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ similar to the ones presented above but with $p_1 = p_2 = 0.95$ were used to simulate populations. The relatively high values of p_1 and p_2, 0.8 and 0.95, ensure that the variability associated with Model ξ is negligible compared with the variability of the two sample designs. Only variability with respect to the design is studied in the simulations. The actual values of N^B and Y^B are the census parameters generated by applying formulas (13) and (16) to the subsets of U^B associated with U^{A1} and U^{A2}, i.e., U_{A1}^B and U_{A2}^B. For the simulations, we used simple random sampling without replacement in U^{A1} and U^{A2} with a sampling fraction of 40%.

The selection probability of a unit k in U^B through files $A1$ ($A2$) is proportional to L_k^{A1} (L_k^{A2}). If L_k^{A1} and L_k^{A2} are independent then the two indirect samples are independent. The standard unweighted Petersen estimator (34) is, in this case, unbiased, and case weighting should make the variance of $\hat{Y}_{CReG}^{B}$ larger than that of the Petersen estimator. The correlation between L_k^{A1} and L_k^{A2} induces heterogeneity in the capture probabilities. The Petersen estimator is then negatively biased (Hook and Regal 1993) and the GWSM corrects this bias. In Tables 1 and 2 the heterogeneity in the capture probabilities is

Table 1. Matrix $\mathbf{M}_1$ for simulating $\left(L_k^{A1}, L_k^{A2}\right)$ for a weak correlation between L_k^{A1} and L_k^{A2}

L_k^{A1}/L_k^{A2}	0	1	2	3	4
0	0.040	0.040	0.040	0.040	0.040
1	0.040	0.100	0.020	0.020	0.020
2	0.040	0.020	0.100	0.020	0.020
3	0.040	0.020	0.020	0.100	0.020
4	0.040	0.020	0.020	0.020	0.100

Table 2. Matrix $\mathbf{M}_2$ *for simulating* $\left(L_k^{A1}, L_k^{A2}\right)$ *for a strong correlation between* L_k^{A1} *and* L_k^{A2}

L_k^{A1}/L_k^{A2}	0	1	2	3	4
0	0.040	0.120	0.040	0.000	0.000
1	0.120	0.136	0.008	0.008	0.008
2	0.040	0.008	0.136	0.008	0.008
3	0.000	0.008	0.008	0.136	0.008
4	0.000	0.008	0.008	0.008	0.136

proportional to the correlation between L_k^{A1} and L_k^{A2}; the simulations compare $\hat{Y}_{Pet}^B$ and $\hat{Y}_{CReG}^B$ under scenarios with low (Table 1) and high (Table 2) heterogeneity.

The results of the simulations are reported in terms of relative bias (*rb*) and root of relative mean squared error (*rrmse*)

$$rb(\hat{\theta}) = \frac{\sum_{i=1}^{T} \hat{\theta}_i / T - \theta}{\theta} \text{ and } rrmse(\hat{\theta}) = \frac{\sqrt{\sum_{i=1}^{T} (\hat{\theta}_i - \theta)^2 / T}}{\theta}$$

where T, the number of simulations, was set at 2,000. These two quantities are reported as percentages in Table 3.

When the correlations between the links with U^{A1} and U^{A2} are weak, the estimators that ignore the links are slightly more biased than the CReG estimators obtained with the GWSM. However, their *rrmse* values are lower when the population sizes are small. In fact, if the links with U^{A1} and U^{A2} are mutually independent, the Petersen estimators are unbiased and more stable than the estimators obtained with the GWSM.

If the correlations between the links with U^{A1} and U^{A2} are strong, the Petersen estimators, which ignore the two sample designs, are biased. In Table 3, the largest bias is about 18%. For all practical purposes, their *rrmse* values are equal to the relative biases. In every case, the estimate of total Y^B is less biased and more stable than the estimate of population size N^B. That is due to the fact that the simulations associate the high values of y_k with high values of L_k^{A1} and L_k^{A2}; as a result, those high values have a high probability of being sampled. Both estimation methods – the CReG method and the Petersen method – benefit from the fact that the high values of y are sampled with high probabilities.

7. Examples

7.1. Estimation of Census Undercount

The CReG estimator can be used in situations where the correspondence between the sampling frames and the target population is not one-to-one. In that case, the units in the two sampling frames are different from the units in the target population. Capture–recapture sampling has to be used when the sampling frames provide only partial coverage of the target population.

We will demonstrate here how the CReG estimator could be used in the Canadian Census of Population. Even though the Census of Population is supposed to be a comprehensive survey, we know that in practice that is not the case. There is no file

Table 3. Relative bias and root of relative mean squared error for four estimators in twelve populations

		Weak correlation ($\mathbf{M}_1$)						Strong correlation ($\mathbf{M}_2$)					
		$p_1 = p_2 = 0.8$			$p_1 = p_2 = 0.95$			$p_1 = p_2 = 0.8$			$p_1 = p_2 = 0.95$		
N^B		**100**	**500**	**2,000**	**100**	**500**	**2,000**	**100**	**500**	**2,000**	**100**	**500**	**2,000**
$\hat{N}^B_{CReG}$	*rb*	1.0	0.2	0.1	1.1	0.1	0.0	1.7	0.2	0.0	0.9	0.3	0.0
	rrmse	13.1	5.9	2.9	10	4.2	2.0	14.5	6.1	3.2	11.1	4.8	2.2
$\hat{N}^B_{Pet}$	*rb*	−2.2	−3.1	−3.3	−3.8	−3.8	−3.2	−14.8	−17.7	−17.8	−8.9	−8.7	−8.2
	rrmse	7.7	4.7	3.7	6.5	4.4	3.4	16.2	17.9	17.9	10.5	9.0	8.2
$\hat{Y}^B_{CReG}$	*rb*	0.7	0.1	0.1	0.8	0.0	0.0	1.2	0.1	0.0	0.8	0.2	0.0
	rrmse	10.4	4.6	2.3	8.6	3.4	1.7	11.2	4.5	2.4	8.9	3.8	1.8
$\hat{Y}^B_{Pet}$	*rb*	−1.8	−2.5	−2.3	−2.7	−2.5	−2.2	−9.6	−10.6	−11.1	−6.2	−5.8	−5.4
	rrmse	6.4	3.7	2.7	5.4	3.1	2.5	11.0	10.7	11.2	7.8	6.1	5.5

containing the entire Canadian population. Note that the coverage problem usually takes the form of an undercount.

Suppose we want to estimate household undercoverage in Canada. Currently, undercoverage is measured at the person level (Statistics Canada 2001). The person undercount is estimated by comparing an outside source of information (a combination of administrative files) with the census, which is a form of capture–recapture using administrative files. The census is the initial sample of persons; the second sample is the outside source of information about persons, which is then matched against the census. The matching of the two sources yields the number of common units. The undercount of persons is then estimated using estimator (2). In practice, a difference estimator is used for the provinces, since administrative files provide excellent coverage. Estimator (2) is used only for the three territories, see Théberge (2008). We will assume here that estimator (2) is used for the whole of Canada.

When it comes to estimating the undercount of households, the problem is complicated by the fact that the sampling units in the initial sample and the second sample are different from the units in the target population. The initial sample and the second sample are sets of persons, while the target population is a set of households. Note that we could attempt to use files of households from the outset and estimate the undercount of households with those files, but that would take a substantial amount of extra work, so we prefer to use the available files of persons. Consequently, we need to use indirect sampling to produce estimates at the household level.

First, we want to reach the target population U^B of Canadian households through the Census of Population. We then get a list U^{A1} of N^{A1} persons in which each person j belongs to a household k. Unfortunately, *that list derived from the census provides only partial coverage*; in other words, it does not contain all Canadian households. Its coverage may be partial because some persons were not counted, or because entire households were not counted. Since that frame is obtained though a census, $\pi_j^{A1} = t_j^{A1} = 1$.

To measure the coverage of the census, we want to reach the target population U^B of Canadian households using a sample of persons. We start with a list U^{A2} of N^{A2} persons; as in the case of the census, *that list provides only partial coverage* of the Canadian population. The list is in fact taken from the previous census and has been updated from various administrative files. From list U^{A2}, we select a sample s^{A2} of n^{A2} persons using a particular sample design. Let π_j^{A2} be the probability of selecting person j, where $\pi_j^{A2} > 0$.

The sample s^{A2} is matched with the list U^{A1} from the census. The $n^{A1,A2}$ persons, present in both s^{A2} and U^{A1}, are assigned to a household from the census. The households of the $n^{A2} - n^{A1,A2}$ persons who were not matched are determined using a field survey. At the end of this process, we have a household identifier for all persons in U^{A1} and s^{A2}. Then the CReG estimator can be used to estimate the total number N^B of households in Canada. Starting from (13), we have

$$\hat{N}^B_{CReG} = \frac{N^B_{A1}\hat{N}^B_{A2}}{\hat{N}^B_{A1,A2}} \tag{35}$$

where N^B_{A1} is the number of U^B households from the census of U^{A1}. The quantity $\hat{N}^B_{A2}$ is determined with the GWSM weights given by (7). We can simplify the calculations by noting that a person j belongs to only one household k. Thus, (8) reduces to

$$\hat{N}^{B}_{A2} = \sum_{j \in s^{A2}} \frac{1}{\pi_j^{A2}} \frac{1}{L_j^{A2}} \tag{36}$$

where L_j^{A2} is the number of persons in the household containing person j.

Lastly, we calculate $\hat{N}^{B}_{A1,A2}$, considering only the households in s^B that are linked to persons in U^{A1} and s^{A2}. We must therefore eliminate the persons, who were not counted in the census, from (36). We get

$$\hat{N}^{B}_{A1,A2} = \sum_{j \in s^{A2}} \frac{1}{\pi_j^{A2}} \frac{\delta_j^{A1}}{L_j^{A2}} \tag{37}$$

where $\delta_{j'}^{A1} = 1$ if the person $j \in s^{A2}$ was counted in U^{A1}, 0 if not. By combining (36) and (37), we obtain the following CReG estimator

$$\hat{N}^{B}_{CReG} = N^{B}_{A1} \frac{\displaystyle\sum_{j \in s^{A1}} \frac{1}{\pi_j^{A2}} \frac{1}{L_j^{A2}}}{\displaystyle\sum_{j \in s^{A1}} \frac{1}{\pi_j^{A2}} \frac{\delta_j^{A1}}{L_j^{A2}}} \tag{38}$$

The sampling frames involved in the construction of this estimator are represented in Figure 2.

Using estimator (38), we are able to estimate the total number N^B of Canadian households, despite the undercoverage in the Census of Population. This estimator is asymptotically unbiased as long as the assumption of independence underlying the

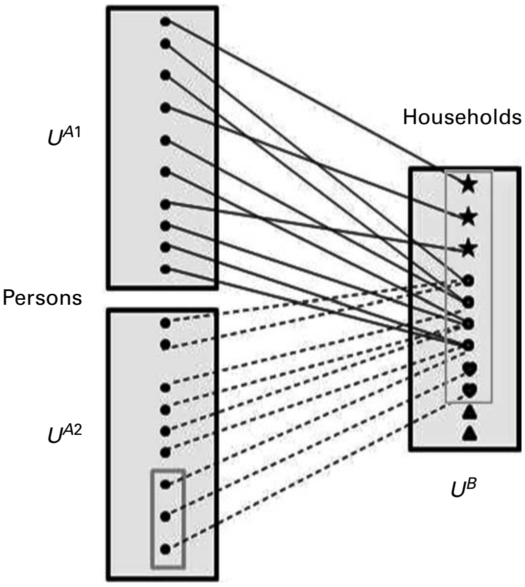

Fig. 2. Census where samples are represented by grey rectangles

Petersen estimator is met; enumeration of a household by the census has to be independent of the coverage of the household by the population U^{A2}.

It should be noted that the indirect sampling operation considered in this section assumed complete response for both the census and administrative files. That is, no nonresponse occurred during the census collection process. For the administrative files, this means that all available records are usable, for instance they all pass edits rules. In practice, adjustments for nonresponse are performed, which mainly consist in adjusting the sampling weights entering in estimator (38).

7.2. Statistics for Tea Vendors

The CReG estimator can be used in a number of situations where, as in the classical case, there is no sampling frame for the target population, and where the available frames are different from the target population and provide only partial coverage. This is often the case in measuring underground economies or economies consisting of micro-businesses (e.g., shoe shiners, itinerant vendors).

We will present a potential application of the CReG estimator with the following example. (This example is inspired by the situation of young tea vendors in the Gaza Strip, which was the subject of a television news story.) Suppose we want to estimate the income of tea vendors in a developing country. Many tea vendors are children, and they make a significant contribution to family income by selling tea to workers and professionals in various parts of the city. They leave home in the morning with a thermos of hot tea and try to sell the tea to people in selected locations: doctors in hospitals, workers on job sites, and so on. They generally keep to the same locations, and customers expect to see their young tea vendor there each day. Of course, the business is viewed as illegal not only because it is a source of unreported income but also because it "employs" children, who should be in school.

We want to estimate the number N^B of child tea vendors and their total daily income Y^B. Obviously, there is no sampling frame for the target population U^B of child tea vendors. We must therefore use indirect sampling.

We can attempt to reach target population U^B by surveying dwellings inhabited by the general population. We can use either a list of addresses or an area frame with a multi-stage sample design, in which dwellings are the final sampling units. For the sake of simplicity, we will assume here that we have an incomplete list U^{A1} of N^{A1} dwellings; in other words, the list does not contain all the dwellings in the city. To survey the target population U^B, we decide to take a stratified sample of dwellings from the list U^{A1}. We stratify into H strata, and stratum h contains N_h^{A1} dwellings; we select n_h^{A1} dwellings in each stratum h by simple random sampling (SRS). In each dwelling j selected from stratum h, we count the set U_{hj}^{A1} of M_{hj}^{A1} tea vendors and measure the daily income y_k of each tea vendor k, $k = 1 \ldots, M_{hj}^{A1}$. Note that because the population of tea vendors is small, using a stratified SRS design might be considered inefficient. It is quite possible to have $M_{hj}^{A1} = 0$, for example. In practice, however, the survey of tea vendors can be part of a much larger survey, such as a labour force survey. Tea vendors are identified through the larger survey, and the actual survey of tea vendors is simply a by-product of the larger survey. This approach is similar to the "1-2-3" surveys described by Bagayogo et al. (2007).

We can also attempt to reach the target population U^B of child tea vendors through their customers. We prepare a list of locations (hospitals, job sites, etc.) that are potential points of sale so that we can survey the child tea vendors' customers from a sample of locations. That gives us a list U^{A2} of N^{A2} locations, but it seems very likely that *this list provides only partial coverage* of the locations where the vendors do business. Because there is a better chance of finding tea vendors in locations frequented by many people, we may decide to sample the locations proportional to their size. Let x_j be the number of people working in location j of U^{A2}. We then define $\pi_j^{A2} = n^{A2} x_j / X^{A2}$, where n^{A2} is the sample size, $X^{A2} = \sum_{j=1}^{N^{A2}} x_j$ and it is assumed that $\pi_j^{A2} \leq 1$ for all locations j. We decide to select the sample s^{A2} using a Poisson design with probabilities π_j^{A2}. In each location j selected, we count the set U_j^{A2} of M_j^{A2} tea vendors and measure the daily income y_k of each tea vendor k, $k = 1 \ldots, M_j^{A2}$ at location j. Again, it is quite possible to have $M_j^{A2} = 0$.

When we survey the tea vendors identified in the sample of dwellings, we make sure to ask them in which public locations they work. We do the same with the tea vendors identified through the sample of public locations. This serves to determine the links $l_{j,k}$ and the total number of links L_k^B needed for the GWSM.

The sampling frames for taking an indirect sample of tea vendors are illustrated in Figure 3.

To estimate Y^B, we use the CReG estimator given by (16). In this case, the component $\hat{Y}_{A1}^B$ of the estimator becomes

$$\hat{Y}_{A1}^B = \sum_{k \in s^{B1}} w_k^{A1} y_k = \sum_{h=1}^{H} \frac{N_h^{A1}}{n_h^{A1}} \sum_{k \in s_h^{B1}} Z_{hj}^{A1} \tag{39}$$

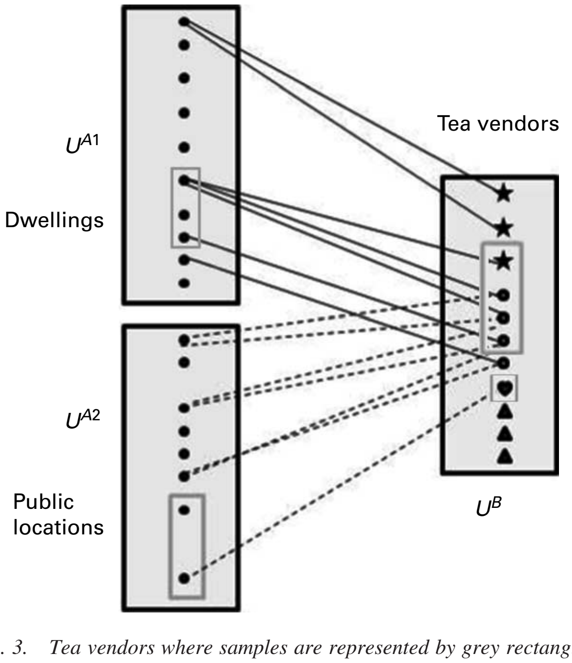

Fig. 3. Tea vendors where samples are represented by grey rectangles

where $Z_{hj}^{A1} = \sum_{k \in U_{hj}^{A1}} y_k$, and s^{B1} denotes the sample of tea vendors obtained from the dwelling frame and s_h^{B1} the sample obtained in stratum h. Note that since a child tea vendor is likely to reside in only one dwelling, we have $L_k^{A1} = 1$ for all $k \in U_{A1}^B$.

The component $\hat{Y}_{A2}^B$ of estimator (16) here becomes

$$\hat{Y}_{A2}^B = \sum_{k \in s^{B2}} w_k^{A2} y_k = \sum_{j \in s^{A2}} \frac{Z_j^{A2}}{\pi_j^{A2}} \tag{40}$$

where $Z_j^{A2} = \sum_{k \in U_j^{A2}} y_k / L_k^{A2}$, with L_k^{A2} representing the number of public locations worked by vendor k.

As for the $\hat{Y}_{A1,A2}^B$ component, we have

$$\begin{aligned}\hat{Y}_{A1,A2}^B &= \sum_{k \in s^{B1} \cap s^{B2}} \left(\frac{1}{L_k^{A1}} \sum_{j=1}^{N^{A1}} l_{j,k}^{A1} \frac{t_j^{A1}}{\pi_j^{A1}} \right) \cdot \left(\frac{1}{L_k^{A2}} \sum_{j'=1}^{N^{A2}} l_{j,k}^{A2} \frac{t_{j'}^{A2}}{\pi_{j'}^{A1}} \right) y_k \\ &= \sum_{j=1}^{N^{A1}} \frac{t_j^{A1}}{\pi_j^{A1}} \sum_{j'=1}^{N^{A2}} \frac{t_{j'}^{A2}}{\pi_{j'}^{A2}} \sum_{k=1}^{n_{A1,A2}^B} l_{j,k}^{A1} l_{j',k}^{A2} \frac{y_k}{L_k^{A2}} = \sum_{h=1}^{H} \frac{N_h^{A1}}{n_h^{A1}} \sum_{j=1}^{n_h^{A1}} \sum_{j'=1}^{\tilde{n}^{A2}} \frac{Z_{j,j'}^{A1,A2}}{\pi_{j'}^{A2}}\end{aligned} \tag{41}$$

where $Z_{j,j'}^{A1,A2} = \sum_{k \in U_{j,j'}^{A1,A2}} y_k / L_k^{A2}$ and $U_{j,j'}^{A1,A2}$ is the set of $M_{j,j'}^{A1,A2}$ tea vendors who live in dwelling j of U^{A1} and sell tea at public location j' of U^{A2}.

Using CReG estimator (16) with components (39), (40) and (41), we can estimate the total income Y^B of child tea vendors (and their number by setting $y_k = 1$), even if the size N^B of the target population is unknown at the outset. This will work as long as the coverage of the tea vendors by the two frames, the households and the public locations are independent.

8. Development of An Estimator by Direct Application of the GWSM

In Section 1 we mentioned that capture–recapture sampling can be associated with the context of sampling with multiple frames. Indirect sampling and the GWSM can also be used in such a context. Examples of such applications of the GWSM are provided in Ardilly and Le Blanc (1999) and Deville and Maumy-Bertrand (2006). Mecatti (2007) proposed a solution similar to the GWSM with a method based on multiplicity, that is, the number of times a unit appears in the various sampling frames.

In the context of multiple frames, population U^A from which the sample is taken is actually constructed from populations U^{A1}, $U^{A2}, \ldots, U^{AQ}$, which are not necessarily exclusive. We have $\cup_{q=1}^{Q} U^{Aq} = U^A$, but $\sum_{q=1}^{Q} N^{Aq} \geq N^A$. Assuming that $Q = 2$, we have two samples s^{A1} and s^{A2} of n^{A1} and n^{A2} units selected from the populations U^{A1} and U^{A2}, respectively and $s^A = s^{A1} \cup s^{A2}$. To estimate the total Y^B for population U^B, we use estimator (6). To estimate the total Y^B for population U^B, we use estimator (6), using a weight w_k assigned to each unit k in $\hat{Y}^B$

$$w_k = \frac{1}{L_k^B} \left[\sum_{j=1}^{N^{A1}} l_{j,k}^{A1} \frac{t_j^{A1}}{\pi_j^{A1}} + \sum_{j=1}^{N^{A2}} l_{j,k}^{A2} \frac{t_j^{A2}}{\pi_j^{A2}} \right] \tag{42}$$

where $L_k^B = \sum_{j=1}^{N^{A1}} l_{j,k}^{A1} + \sum_{j=1}^{N^{A2}} l_{j,k}^{A2}$, $t_j^{A1} = 1$ if $j \in s^{A1}$ and 0 if not, and similarly for t_j^{A2}.

Since capture–recapture sampling is associated with the context of multiple frames, we might have considered applying the GWSM directly in the same way as in Section 3.1. What follows shows that the solution obtained in that case is problematic.

First, we consider the simple case in which population U^{A1}, population U^{A2} and target population U^B are identical. This is the most common case in the application of capture–recapture sampling. We begin by selecting an initial sample s^{A1} of size n^{A1} from population U^{A1} of size N^B (unknown). The second sample s^{A2} of n^{A2} units is selected from population U^{A2}, which in this case is identical to population U^{A1}. We want to estimate the size N^B of target population U^B, which is again identical to population U^{A1}. This indirect sampling process is illustrated schematically in Figure 4 below.

Applying the GWSM directly, we obtain the following result

$$\hat{N}^B = \sum_{k \in s^A} w_k \tag{43}$$

where weight w_k is given by (42). Since there are exactly two links for each unit k of U^B, we have $L_k^B = 2$ and

$$w_k = \frac{1}{2}\left(\frac{t_j^{A1}}{\pi_j^{A1}} + \frac{t_{j'}^{A2}}{\pi_{j'}^{A2}}\right),$$

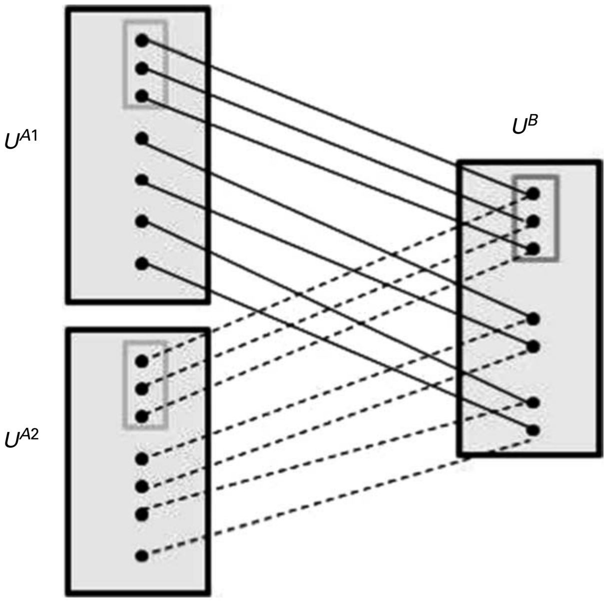

Fig. 4. Capture–recapture sampling with U^{A1}, U^{A2} and U^B identical where samples are represented by grey rectangles

where $j \in U^{A1}$ and $j' \in U^{A2}$ both have links with k. Note that because $U^{A1} = U^{A2} = U^B$, indexes k, j and j' are interchangeable, and we can write

$$\hat{N}^B = \frac{1}{2}\sum_{k \in s^A}\left(\frac{t_k^{A1}}{\pi_k^{A1}} + \frac{t_k^{A2}}{\pi_k^{A2}}\right) \tag{44}$$

This is an estimator for the size of the target population U^B.

Unfortunately, there is a serious problem with estimator (44). Since both samples s^{A1} and s^{A2} are from the same frame, which is also the same as the target population, we have $U^{A1} = U^{A2} = U^B$ and $N^{A1} = N^{A2} = N^B$. The selection probabilities π_k^{A1} and π_k^{A2} are generally associated with the size of their respective population (or a quantity related to that population) and therefore depend on N^{A1}, N^{A2}, and N^B. For example, with simple random sampling, we have $\pi_k^{A1} = n^{A1}/N^{A1}$, and consequently in this case $\pi_k^{A1} = n^{A1}/N^B$. However, it is precisely N^B that we are trying to estimate, so we do not have the selection probabilities π_k^{A1} and π_k^{A2} needed to use estimator (44)! In this context, estimator (44) is not very useful. For it to be usable in practice, the selection probabilities π_k^{A1} and π_k^{A2} *should not* be directly involved in the estimator; that is the case with estimators (13) and (16) when the selection probabilities are equal within each sampling frame U^{A1} and U^{A2}.

The same problem arises for the administrative files, though in a different way. As previously mentioned, in this context, the target population U^B is only partially covered by a set of sampling frames (or administrative files) U^{A1} and U^{A2}. We are interested in the case where a census of U^{A1} and U^{A2} is impossible. We therefore use two samples s^{A1} and s^{A1} drawn from administrative files U^{A2} and U^{A2}, respectively. With sample s^{A1} from the first file, we can produce an unbiased estimate $\hat{N}^{A1}$ of N^{A1} using the Horvitz-Thompson estimator $\hat{N}^{A1} = \sum_{k \in s^{A1}} 1/\pi_k^{A1}$. With s^{A2}, we can similarly obtain an unbiased estimate $\hat{N}^{A2}$ of N^{A2}. Unfortunately, since each file provides only a partial coverage of the target population, neither of the estimates $\hat{N}^{A1}$ and $\hat{N}^{A2}$ provides an unbiased estimate of the size N^B of the target population U^B.

Another potential solution to the administrative files problem is to estimate N^B using the GWSM with the weights given by (42) with selection probabilities π_k^{A1} and π_k^{A2}. Thus, we have

$$\begin{aligned}\hat{N}^B &= \sum_{k=1}^{N^B}\frac{1}{L_k^B}\left[\sum_{j=1}^{N^{A1}} l_{j,k}^{A1}\,\frac{t_j^{A1}}{\pi_j^{A1}} + \sum_{j=1}^{N^{A2}} l_{j,k}^{A2}\,\frac{t_j^{A2}}{\pi_j^{A2}}\right] \\ &= \sum_{j=1}^{N^{A1}}\frac{t_j^{A1}}{\pi_j^{A1}}\sum_{k=1}^{N^B}\frac{l_{j,k}^{A1}}{L_k^B} + \sum_{j=1}^{N^{A2}}\frac{t_j^{A2}}{\pi_j^{A2}}\sum_{k=1}^{N^B}\frac{l_{j,k}^{A2}}{L_k^B}\end{aligned} \tag{45}$$

As previously discussed, for the GWSM to be unbiased, we must have $L_k^B > 0$; in other words, each unit k of U^B must have at least one link with a unit j of U^{A1} or U^{A2}. Because the target population U^B is only partially covered by the administrative files U^{A1} and U^{A2}, there are some units k of U^B for which $L_k^B = 0$. In fact, estimator (45) is unbiased for the estimation of N^B_{A1+A2}, the size of the population $U^B_{A1+A2} = \left\{k \in U^B | \exists j \in U^{A1} \cup U^{A2}, l^{A1}_{j,k} \text{ or } l^{A2}_{j,k} \neq 0\right\}$, i.e., the subpopulation of U^B whose units have at least one link with U^{A1} or U^{A2}. We conclude from this that for administrative files, estimator (45) based on the GWSM cannot provide unbiased estimates of N^B.

The CReG estimator was constructed to address the problems encountered in direct application of the GWSM. It provides a solution to the bias issues highlighted in this section.

Appendix A

We want to show that (15) is unbiased for the estimation of $Y^B_{A1,A2}$. Thus, we have

$$\hat{Y}^B_{A1,A2} = \sum_{k=1}^{N^B_{A1,A2}} \left(\frac{1}{L^{A1}_k} \sum_{j=1}^{N^{A1}} l^{A1}_{j,k} \frac{t^{A1}_j}{\pi^{A1}_j} \right) \cdot \left(\frac{1}{L^{A2}_k} \sum_{j'=1}^{N^{A2}} l^{A2}_{j',k} \frac{t^{A2}_{j'}}{\pi^{A2}_{j'}} \right) y_k$$

Since the samples from U^{A1} and U^{A2} are independent, the expected value is

$$E_p\left(\hat{Y}^B_{A1,A2}\right) = \sum_{k=1}^{N^B_{A1,A2}} \left(\frac{1}{L^{A1}_k} \sum_{j=1}^{N^{A1}} l^{A1}_{j,k} \frac{E_p\left(t^{A1}_j\right)}{\pi^{A1}_j} \right) \cdot \left(\frac{1}{L^{A2}_k} \sum_{j'=1}^{N^{A2}} l^{A2}_{j',k} \frac{E_p\left(t^{A2}_{j'}\right)}{\pi^{A2}_{j'}} \right) y_k$$

$$= \sum_{k=1}^{N^B_{A1,A2}} \left(\frac{1}{L^{A1}_k} \sum_{j=1}^{N^{A1}} l^{A1}_{j,k} \right) \cdot \left(\frac{1}{L^{A2}_k} \sum_{j'=1}^{N^{A2}} l^{A2}_{j',k} \right) y_k \quad = \sum_{k=1}^{N^B_{A1,A2}} (1) \cdot (1)\, y_k = Y^B_{A1,A2}$$

We therefore have $E_p\left(\hat{Y}^B_{A1,A2}\right) = Y^B_{A1,A2}$.

Appendix B

Calculation of the variance $V_p\left(\hat{Y}^B_{A1,A2}\right)$:

$$V_p\left(\hat{Y}^B_{A1,A2}\right) = E_p\left(\hat{Y}^B_{A1,A2} \times \hat{Y}^B_{A1,A2}\right) - \left(Y^B_{A1,A2}\right)^2$$

$$= E_p\left(\sum_{k=1}^{N^B_{A1,A2}} \left(\frac{1}{L^{A1}_k} \sum_{j=1}^{N^{A1}} l^{A1}_{j,k} \frac{t^{A1}_j}{\pi^{A1}_j} \right) \left(\frac{1}{L^{A2}_k} \sum_{j''=1}^{N^{A2}} l^{A2}_{j'',k} \frac{t^{A1}_{j''}}{\pi^{A1}_{j''}} \right) y_k \right.$$

$$\left. \times \sum_{k'=1}^{N^B_{A1,A2}} \left(\frac{1}{L^{A1}_{k'}} \sum_{j'=1}^{N^{A1}} l^{A1}_{j',k'} \frac{t^{A1}_{j'}}{\pi^{A1}_{j'}} \right) \cdot \left(\frac{1}{L^{A2}_{k'}} \sum_{j'''=1}^{N^{A2}} l^{A2}_{j''',k'} \frac{t^{A2}_{j'''}}{\pi^{A2}_{j'''}} \right) y_{k'} \right) - \left(Y^B_{A1,A2}\right)^2$$

$$
\begin{aligned}
&=E_p\left(\sum_{j=1}^{N^{A1}}\frac{t_j^{A1}}{\pi_j^{A1}}\sum_{j''=1}^{N^{A2}}\frac{t_{j''}^{A2}}{\pi_{j''}^{A2}}\left(\sum_{k=1}^{N_{A1,A2}^{B}} l_{j,k}^{A1} l_{j",k}^{A2}\frac{y_k}{L_k^{A1}L_k^{A2}}\right)\right.\\
&\qquad\left.\times\sum_{j'=1}^{N^{A1}}\frac{t_{j'}^{A1}}{\pi_{j'}^{A1}}\sum_{j'''=1}^{N^{A2}}\frac{t_{j'''}^{A2}}{\pi_{j'''}^{A2}}\left(\sum_{k'=1}^{N_{A1,A2}^{B}} l_{j',k'}^{A1} l_{j''',k'}^{A2}\frac{y_{k'}}{L_{k'}^{A1}L_{k'}^{A2}}\right)\right)-\left(Y_{A1,A2}^{B}\right)^2\\
&=E_p\left(\sum_{j=1}^{N^{A1}}\frac{t_j^{A1}}{\pi_j^{A1}}\sum_{j''=1}^{N^{A2}}\frac{t_{j''}^{A2}}{\pi_{j''}^{A2}}Z_{jj''}^{A1,A2}\times\sum_{j'=1}^{N^{A1}}\frac{t_{j'}^{A1}}{\pi_{j'}^{A1}}\sum_{j'''=1}^{N^{A2}}\frac{t_{j'''}^{A2}}{\pi_{j'''}^{A2}}Z_{j'j'''}^{A1,A2}\right)-\left(Y_{A1,A2}^{B}\right)^2\\
&=E_p\left(\sum_{j=1}^{N^{A1}}\sum_{j'=1}^{N^{A1}}\frac{t_j^{A1}t_{j'}^{A1}}{\pi_j^{A1}\pi_{j'}^{A1}}\sum_{j''=1}^{N^{A2}}\sum_{j'''=1}^{N^{A2}}\frac{t_{j''}^{A2}t_{j'''}^{A2}}{\pi_{j''}^{A2}\pi_{j'''}^{A2}}Z_{jj''}^{A1,A2}Z_{j'j'''}^{A1,A2}\right)-\left(Y_{A1,A2}^{B}\right)^2\\
&=\sum_{j=1}^{N^{A1}}\sum_{j'=1}^{N^{A1}}\frac{E_p\left(t_j^{A1}t_{j'}^{A1}\right)}{\pi_j^{A1}\pi_{j'}^{A1}}\sum_{j''=1}^{N^{A2}}\sum_{j'''=1}^{N^{A2}}\frac{E_p\left(t_{j''}^{A2}t_{j'''}^{A2}\right)}{\pi_{j''}^{A2}\pi_{j'''}^{A2}}Z_{jj''}^{A1,A2}Z_{j'j'''}^{A1,A2}-\left(Y_{A1,A2}^{B}\right)^2\\
&=\sum_{j=1}^{N^{A1}}\sum_{j'=1}^{N^{A1}}\frac{\pi_{jj'}^{A1}}{\pi_j^{A1}\pi_{j'}^{A1}}\sum_{j''=1}^{N^{A2}}\sum_{j'''=1}^{N^{A2}}\frac{\pi_{j''j'''}^{A2}}{\pi_{j''}^{A2}\pi_{j'''}^{A2}}Z_{jj''}^{A1,A2}Z_{j'j'''}^{A1,A2}-\left(Y_{A1,A2}^{B}\right)^2
\end{aligned}
$$

where $E_p\left(t_j^{A1}t_{j''}^{A1}\right)=\pi_{jj''}^{A1}=P(j\in s^{A1},j''\in s^{A1})$ and $E_p\left(t_{j'}^{A2}t_{j'''}^{A2}\right)=\pi_{j'j'''}^{A2}=P(j'\in s^{A2}, j'''\in s^{A2})$.

Calculation of covariances $Cov_p\left(\hat{Y}_{A1}^{B},\hat{Y}_{A1,A2}^{B}\right)$ **and** $Cov_p\left(\hat{Y}_{A2}^{B},\hat{Y}_{A1,A2}^{B}\right)$:

$$
\begin{aligned}
&Cov_p\left(\hat{Y}_{A1}^{B},\hat{Y}_{A1,A2}^{B}\right)\\
&=E_p\left(\hat{Y}_{A1}^{B}\times\hat{Y}_{A1,A2}^{B}\right)-Y_{A1}^{B}Y_{A1,A2}^{B}\\
&=E_p\left(\sum_{k=1}^{N_{A1}^{B}}\left(\frac{1}{L_k^{A1}}\sum_{j=1}^{N^{A1}} l_{j,k}^{A1}\frac{t_j^{A1}}{\pi_j^{A1}}\right)y_k\times\sum_{k'=1}^{N_{A1,A2}^{B}}\left(\frac{1}{L_{k'}^{A1}}\sum_{j'=1}^{N^{A1}} l_{j',k'}^{A1}\frac{t_{j'}^{A1}}{\pi_{j'}^{A1}}\right)\cdot\left(\frac{1}{L_{k'}^{A2}}\sum_{j''=1}^{N^{A2}} l_{j'',k'}^{A2}\frac{t_{j''}^{A2}}{\pi_{j''}^{A2}}\right)y_{k'}\right)\\
&\quad-Y_{A1}^{B}Y_{A1,A2}^{B}\\
&=E_p\left(\sum_{j=1}^{N^{A1}}\frac{t_j^{A1}}{\pi_j^{A1}}\left(\sum_{k=1}^{N_{A1}^{B}} l_{j,k}^{A1}\frac{y_k}{L_k^{A1}}\right)\times\sum_{j'=1}^{N^{A1}}\frac{t_{j'}^{A1}}{\pi_{j'}^{A1}}\sum_{j''=1}^{N^{A2}}\frac{t_{j''}^{A2}}{\pi_{j''}^{A2}}\left(\sum_{k'=1}^{N_{A1,A2}^{B}} l_{j',k'}^{A1} l_{j'',k'}^{A2}\frac{y_{k'}}{L_{k'}^{A1}L_{k'}^{A2}}\right)\right)\\
&\quad-Y_{A1}^{B}Y_{A1,A2}^{B}
\end{aligned}
$$

$$= E_p\left(\sum_{j=1}^{N^{A1}} \frac{t_j^{A1}}{\pi_j^{A1}} Z_j^{A1} \times \sum_{j'=1}^{N^{A1}} \frac{t_{j'}^{A1}}{\pi_{j'}^{A1}} \sum_{j''=1}^{N^{A2}} \frac{t_{j''}^{A2}}{\pi_{j''}^{A2}} Z_{j'j''}^{A1,A2}\right) - Y_{A1}^B Y_{A1,A2}^B$$

$$= E_p\left(\sum_{j=1}^{N^{A1}} \sum_{j'=1}^{N^{A1}} \frac{t_j^{A1} t_{j'}^{A1}}{\pi_j^{A1} \pi_{j'}^{A1}} Z_j^{A1} \sum_{j''=1}^{N^{A2}} \frac{t_{j''}^{A2}}{\pi_{j''}^{A2}} Z_{j'j''}^{A1,A2}\right) - Y_{A1}^B Y_{A1,A2}^B$$

$$= \sum_{j=1}^{N^{A1}} \sum_{j'=1}^{N^{A1}} \frac{E_p\left(t_j^{A1} t_{j'}^{A1}\right)}{\pi_j^{A1} \pi_{j'}^{A1}} Z_j^{A1} \sum_{j''=1}^{N^{A2}} \frac{E_p\left(t_{j''}^{A2}\right)}{\pi_{j''}^{A2}} Z_{j'j''}^{A1,A2} - Y_{A1}^B Y_{A1,A2}^B$$

$$= \sum_{j=1}^{N^{A1}} \sum_{j'=1}^{N^{A1}} \frac{\pi_{jj'}^{A1}}{\pi_j^{A1} \pi_{j'}^{A1}} Z_j^{A1} \sum_{j''=1}^{N^{A2}} Z_{j'j''}^{A1,A2} - Y_{A1}^B Y_{A1,A2}^B$$

where $E_p\left(t_j^{A1} t_{j'}^{A1}\right) = \pi_{jj'}^{A1} = P(j \in s^{A1}, j' \in s^{A1})$ and $E_p\left(t_{j''}^{A2}\right) = \pi_{j''}^{A2} = P(j'' \in s^{A2})$.

Now, it is simple to verify that $\sum_{j''=1}^{N^{A2}} Z_{j'j''}^{A1A2} = Z_{j'}^{A1A2}$, where $Z_{j'}^{A1,A2} = \sum_{k=1}^{N_{A1,A2}^B} l_{j',k}^{A1} y_k / L_k^{A1}$. Thus,

$$Cov_p\left(\hat{Y}_{A1}^B, \hat{Y}_{A1,A2}^B\right) = \sum_{j=1}^{N^{A1}} \sum_{j'=1}^{N^{A1}} \frac{\pi_{jj'}^{A1}}{\pi_j^{A1} \pi_{j'}^{A1}} Z_j^{A1} Z_{j'}^{A1,A2} - Y_{A1}^B Y_{A1,A2}^B$$

$$= \sum_{j=1}^{N^{A1}} \sum_{j'=1}^{N^{A1}} \frac{\pi_{jj'}^{A1} - \pi_j^{A1} \pi_{j'}^{A1}}{\pi_j^{A1} \pi_{j'}^{A1}} Z_j^{A1} Z_{j'}^{A1,A2}$$

Similarly, we obtain

$$Cov_p\left(\hat{Y}_{A2}^B, \hat{Y}_{A1,A2}^B\right) = \sum_{j=1}^{N^{A2}} \sum_{j'=1}^{N^{A2}} \frac{\pi_{jj'}^{A2}}{\pi_j^{A2} \pi_{j'}^{A2}} Z_j^{A2} Z_{j'}^{A1,A2} - Y_{A2}^B Y_{A1,A2}^B$$

$$= \sum_{j=1}^{N^{A2}} \sum_{j'=1}^{N^{A2}} \frac{\pi_{jj'}^{A2} - \pi_j^{A2} \pi_{j'}^{A2}}{\pi_j^{A2} \pi_{j'}^{A2}} Z_j^{A2} Z_{j'}^{A1,A2}$$

9. References

Ardilly, P. and LeBlanc, P. (1999). Enquête auprès des personnes sans domicile: éléments techniques sur l'échantillonnage et le calcul de pondérations individuelles – Une application de la méthode du partage des poids. Document de travail de l'INSEE, No. F9903. [In French]

Alho, J.M. (1990). Logistic Regression in Capture–Recapture Models. Biometrics, 46, 623–635.

Bagayogo, A., Ouedraogo, E., and Vescovo, A. (2007). Effet du plan de sondage dans les enquêtes en phases: les enquêtes 1-2-3 en Afrique de l'Ouest. Article présenté au cinquième Colloque francophone sur les sondages, Marseille, 5 au 7 novembre 2007. [In French].

Chapman, D.G. (1951). Some Properties of the Hypergeometric Distribution with Applications to Zoological Censuses. University of California Publications in Statistics, 1, 131–160.

Chen, S.X. and Lloyd, S.J. (2000). A Nonparametric Approach to the Analysis of Two-stage Mark-recapture Experiments. Biometrika, 87, 633–649.

Deville, J.-C. and Maumy-Bertrand, M. (2006). Extension of the Indirect Sampling Method and Its Application to Tourism. Survey Methodology, 32, 177–185.

Ernst, L. (1989). Weighting Issues for Longitudinal Household and Family Estimates. Panel Surveys, D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh (eds). New York: John Wiley and Sons, 135–159.

Hook, E.B. and Regal, R.R. (1993). Effect of Variation in Probability of Ascertainment by Sources (Variable Catchability) upon Capture-Recapture Estimates of Prevalence. American Journal of Epidemiology, 137, 1148–1166.

Laplace, P.S. (1786). Sur les naissances, les mariages et les morts. Histoire de l'Académie Royale des Sciences. Année 1783, p. 693. [In French]

Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households using the Weight Share Method. Survey Methodology, 21, 25–32.

Lavallée, P. (2002). Le Sondage indirect, ou la méthode généralisée du partage des poids. Éditions de l'Université de Bruxelles, Belgique, Éditions Ellipse, France. [In French]

Lavallée, P. (2007). Indirect Sampling. New York: Springer.

Le Cren, E.D. (1965). A Note on the History of Mark-Recapture Population Estimates. Journal of Animal Ecology, 34, 453–454.

Mecatti, F. (2007). A Single Frame Multiplicity Estimator for Multiple Frame Surveys. Survey Methodology, 33, 151–158.

Otis, D.L., Burnham, K.P., White, G.C., and Anderson, D.R. (1978). Statistical Inference from Capture Data on Closed Animal Populations. Wildlife Monograph, 62, 1–135.

Plante, N., Rivest, L.-P., and Tremblay, R. (1998). Stratified Capture Recapture Estimation of the Size of a Closed Population. Biometrics, 54, 47–60.

Rivest, L.-P., Potvin, F., Crepeau, H., and Daigle, G. (1995). Statistical Methods for Aerial Surveys using the Double-count Techniques to Correct Visibility Bias. Biometrics, 51, 461–470.

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

Seber, G.A.F. (1970). The Effects of Trap Response on Tag Recapture Estimates. Biometrics, 26, 13–22.

Seber, G.A.F. (1982). The Estimation of Animal Abundance and Related Parameters, (Second Edition). London: Griffin.

Sekar, C.C. and Deming, W.E. (1949). On a Method of Estimating Birth and Death Rates and the Extent of Registration. Journal of the American Statistical Association, 44, 101–115.

Statistics Canada (2001). Coverage, 2001 Census. Technical Report, Catalogue Number 92-394-X1E, Ottawa.

Théberge, A. (2008). Estimation des ménages et des familles. Document de travail de la Division des méthodes d'enquêtes sociales, Statistics Canada, December. [In French]

Thompson, S.K. (2002). Sampling, (Second Edition). New York: John Wiley and Sons.

Thompson, S.K. and Seber, G.A. (1996). Adaptive Sampling. New York: John Wiley and Sons.

Yates, F. and Grundy, P.M. (1953). Selection without Replacement from Within Strata with Probability Proportional to Size. Journal of the Royal Statistical Society, Series B, 15, 235–261.

Received February 2010
Revised March 2011