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

Survey Methodology

Estimation methods on multiple sampling frames in two-stage sampling designs

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Canada

by Guillaume Chauvet and Guylène Tandeau de Marsac

Release date: December 19, 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by "Key resource" > "Publications."

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. 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.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Published by authority of the Minister responsible for Statistics Canada

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 co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2014
Vol. 40, No. 2, p. 335-346
Statistics Canada, Catalogue No. 12-001-X

Estimation methods on multiple sampling frames in two-stage sampling designs

Guillaume Chauvet and Guylène Tandeau de Marsac[1]

Abstract

When studying a finite population, it is sometimes necessary to select samples from several sampling frames in order to represent all individuals. Here we are interested in the scenario where two samples are selected using a two-stage design, with common first-stage selection. We apply the Hartley (1962), Bankier (1986) and Kalton and Anderson (1986) methods, and we show that these methods can be applied conditional on first-stage selection. We also compare the performance of several estimators as part of a simulation study. Our results suggest that the estimator should be chosen carefully when there are multiple sampling frames, and that a simple estimator is sometimes preferable, even if it uses only part of the information collected.

Key Words: Expansion survey; Hansen-Hurwitz estimator; Horvitz-Thompson estimator; Two-stage sampling.

1 Introduction

When studying a finite population, sometimes no sampling frame covers that population completely, and it is necessary to select samples from two or more sampling frames in order to represent all individuals. Many methods of estimation on multiple sampling frames have been proposed to pool these samples (Hartley 1962; Bankier 1986; Kalton and Anderson 1986; Mecatti 2007; Rao and Wu 2010); see also the review articles by Lohr (2009, 2011) and the referenced articles for a complete picture. Note that the Mecatti method (2007) is inspired by the work of Lavallée (2002, 2007) on the Generalized Weight Share Method. In Section 2, we present different estimation methods for multiple sampling frames.

In Section 3, we are interested in the scenario where two samples are selected using a two-stage design, with common first-stage selection. This framework corresponds to INSEE expansion surveys: an initial sample of dwellings is selected from the communes of the master sample (Bourdalle, Christine and Wilms 2000), and a second sample is selected and surveyed from the communes of the same master sample to target a specific subpopulation. We have two survey measurements from two independent samples at the second stage of the design. We apply estimation methods to multiple sampling frames to pool these two samples. We show that the estimators examined can in this case be calculated conditional on the first stage of selection, which simplifies calculation particularly for Hartley’s optimal estimator (1962). In Section 4, we compare the performance of these estimators as part of a simulation study. We present our conclusion in Section 5.

[1] Guillaume Chauvet, ENSAI (CREST), Ker Lann Campus, Bruz, France. Email: chauvet@ensai.fr. Guylène Tandeau de Marsac, INSEE, Regional Direction of Lille, France. Email: guylene.tandeau-de-marsac@insee.fr.

2 Estimation for multiple sampling frames

A finite population U upon which is defined a variable of interest y of value y_k for individual k is considered. If a sample S is selected from U with inclusion probabilities π_k, the estimator $\hat{Y} = \sum_{k \in S} \pi_k^{-1} y_k$ proposed by Narain (1951) and Horvitz and Thompson (1952) is unbiased for total $Y = \sum_{k \in U} y_k$ if all probabilities π_k are strictly positive.

We are interested in the scenario where the population is fully covered by two overlapping sampling frames, U_A and U_B. We used Lohr's (2011) notation, namely $a = U_A \setminus U_B$ the domain covered by U_A only; $b = U_B \setminus U_A$ the domain covered by U_B only; $ab = U_A \cap U_B$ the domain covered both by U_A and U_B. A sample S^A is selected in U_A with inclusion probabilities $\pi_k^A > 0$. For any domain $d \subset U_A$, the sub-total $Y_d = \sum_{k \in d} y_k$ is unbiasedly estimated by $\hat{Y}_d^A = \sum_{k \in S_A} d_k^A y_k 1(k \in d)$ with $d_k^A = \left(\pi_k^A\right)^{-1}$. A sample U_B is selected in S^B with inclusion probabilities $\pi_k^B > 0$. For any domain $d \subset U_B$, the sub-total Y_d is unbiasedly estimated by $\hat{Y}_d^B = \sum_{k \in S_B} d_k^B y_k 1(k \in d)$ with $d_k^B = \left(\pi_k^B\right)^{-1}$. The objective is to combine the samples S^A and S^B to get estimation Y as accurate as possible.

2.1 Hartley estimator

Hartley (1962) proposes the class of unbiased estimators

$$\hat{Y}_\theta = \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1-\theta) \hat{Y}_{ab}^B + \hat{Y}_b^B, \tag{2.1}$$

with θ one parameter to be determined. The choice $\theta = 1/2$ gives samples S^A and S^B the same weight for the estimation on the intersection domain ab. Hartley (1962) proposes choosing the parameter that minimizes the variance of $\hat{Y}_\theta$. This leads to

$$\theta_{opt} = \frac{Cov\left(\hat{Y}_a^A + \hat{Y}_{ab}^B + \hat{Y}_b^B, \hat{Y}_{ab}^B - \hat{Y}_{ab}^A\right)}{V\left(\hat{Y}_{ab}^B - \hat{Y}_{ab}^A\right)}, \tag{2.2}$$

which can be re-expressed as

$$\theta_{opt} = \frac{V\left(\hat{Y}_{ab}^B\right) + Cov\left(\hat{Y}_{ab}^B, \hat{Y}_b^B\right) - Cov\left(\hat{Y}_a^A, \hat{Y}_{ab}^A\right)}{V\left(\hat{Y}_{ab}^A\right) + V\left(\hat{Y}_{ab}^B\right)} \tag{2.3}$$

when the samples S^A and S^B are independent. As noted by Lohr (2007), the optimal coefficient θ_{opt} may not be between 0 and 1 if a covariance term present in (2.3) is large. To simplify, let us assume that $Cov\left(\hat{Y}_{ab}^B, \hat{Y}_b^B\right) = 0$, which is the case if b and ab are used as strata in the selection of S^B. Then $\theta_{opt} > 1$ if and only if $Cov\left(\hat{Y}^A, \hat{Y}_{ab}^A\right) < 0$. When S^A is selected by simple random sampling, this will be the case, for example, if in U_A the low values of the variable y are concentrated in the domain ab.

In practice, the variance and covariance terms are unknown and must be replaced by estimators, which introduces additional variability. Another disadvantage is that the optimal parameter depends on the variable of interest considered. If optimal estimators are calculated for different variables of interest, estimations may be internally inconsistent (Lohr 2011).

2.2 Kalton and Anderson estimator

A more general class of estimators is obtained by noting that total Y can be re-expressed as

$$Y = Y_a + \sum_{k \in ab} \theta_k y_k + \sum_{k \in ab} \left(1 - \theta_k\right) y_k + Y_b,$$

with θ_k a coefficient specific to the individual k. Kalton and Anderson (1986) propose the choice $\theta_k = \left(d_k^A + d_k^B\right)^{-1} d_k^B$, which leads to the estimator

$$\hat{Y}_{KA} = \sum_{k \in S^A} d_k^A m_k^A y_k + \sum_{k \in S^B} d_k^B m_k^B y_k \tag{2.4}$$

with on one hand $m_k^A = 1$ if $k \in a$ and $m_k^A = \theta_k$ if $k \in ab$, and on the other hand $m_k^B = 1$ if $k \in b$ and $m_k^B = 1 - \theta_k$ if $k \in ab$. The estimation weights are the same regardless of the variable of interest, which guarantees internal consistency of the estimations; on the other hand, the Kalton and Anderson estimator is less effective than Hartley's optimal estimator for a given variable of interest. Note that it is a Hansen-Hurwitz (1943) type estimator, which can be re-expressed as $\hat{Y}_{KA} = \sum_{k \in U} \left[W_k / E\left(W_k\right)\right] y_k$ noting $W_k = 1\left(k \in S^A\right) + 1\left(k \in S^B\right)$ the number of times when unit k is selected in the pooled sample $S^A \cup S^B$. In particular this gives $E(W_k) = \pi_k^A + \pi_k^B$.

2.3 Bankier estimator

Bankier (1986) proposes using a Horvitz-Thompson type estimator, calculating the inclusion probabilities in the pooled sample.

$$\pi_k^{HT} \equiv P\left(k \in S^A \cup S^B\right) = \pi_k^A + \pi_k^B - Pr\left(k \in S^A \cap S^B\right).$$

If the samples S^A and S^B are independent, we get $\pi_k^{HT} = \pi_k^A + \pi_k^B - \pi_k^A \pi_k^B$ and the estimator

$$\hat{Y}_{HT} = \sum_{k \in S^A \cup S^B} \frac{y_k}{\pi_k^{HT}} = \sum_{k \in S^A \cap a} \frac{y_k}{\pi_k^A} + \sum_{k \in S^B \cap b} \frac{y_k}{\pi_k^B} + \sum_{k \in \left(S^A \cup S^B\right) \cap ab} \frac{1}{\pi_k^A + \pi_k^B - \pi_k^A \pi_k^B} y_k. \tag{2.5}$$

3 Estimation with common first-stage selection

Here we are interested in the case of two samples selected using a two-stage design, with common first-stage selection. Population U is partitioned to obtain a population $U_I = \left\{u_1, \ldots, u_M\right\}$ of M primary sampling units. In the first stage, a sample S_I of primary sampling units (PSU) is selected, with a selection probability π_{Ii} for a PSU u_i. In the second stage, in each primary sampling unit $u_i \in S_I$, the following is selected: a sample S_i^A in $u_i^A \equiv u_i \cap U_A$, with a (conditional) selection probability $\pi_{k|i}^A > 0$ for

$k \in u_i^A$; a sample S_i^B in $u_i^B \equiv u_i \cap U_B$, with a (conditional) selection probability $\pi_{k|i}^B > 0$ for unit $k \in u_i^B$. We make the following hypotheses, which are common for two-stage selection: the second stage of selection in the primary sampling unit u_i depends only on i; between two primary sampling units $u_i \neq u_j \in S_I$, the samples S_i^A and S_j^A (respectively, S_i^B and S_j^B) are conditionally independent to S_I (property of independence). We also assume that within each primary sampling unit $u_i \in S_I$, the sub-samples S_i^A and S_i^B are conditionally independent to S_I.

For a domain $d_1 \subset U_A$, the sub-total Y_{d_1} is estimated by $\hat{Y}_{d_1}^A = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{d_1,i}^A$ with $d_{Ii} = (\pi_{Ii})^{-1}$ the sampling weight of the primary sampling unit u_i, $\hat{Y}_{d_1,i}^A = \sum_{k \in S_i^A} d_{k|i}^A y_k 1(k \in d_1)$ the estimator of the sub-total $Y_{d_1,i} = \sum_{k \in u_i} y_k 1(k \in d_1)$ over $d_1 \cap u_i$, and $d_{k|i}^A = (\pi_{k|i}^A)^{-1}$ the sampling weight of k in u_i^A. For a domain $d_2 \subset U_B$, the sub-total Y_{d_2} is estimated by $\hat{Y}_{d_2}^B = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{d_2,i}^B$ with $\hat{Y}_{d_2,i}^B = \sum_{k \in S_i^B} d_{k|i}^B y_k 1(k \in d_2)$ the estimator of the sub-total $Y_{d_2,i}$ and $d_{k|i}^B = (\pi_{k|i}^B)^{-1}$ the sampling weight of k in u_i^B. This yields in particular the estimators

$$\hat{Y}_{ab}^A = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{ab,i}^A \text{ where } \hat{Y}_{ab,i}^A = \sum_{k \in S_i^A} d_{k|i}^A y_k 1(k \in ab), \tag{3.1}$$

$$\hat{Y}_{b}^A = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{b,i}^A \text{ where } \hat{Y}_{b,i}^A = \sum_{k \in S_i^A} d_{k|i}^A y_k 1(k \in b), \tag{3.2}$$

$$\hat{Y}_{ab}^B = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{ab,i}^B \text{ where } \hat{Y}_{ab,i}^B = \sum_{k \in S_i^B} d_{k|i}^B y_k 1(k \in ab). \tag{3.3}$$

3.1 Hartley estimator

The Hartley estimator given in (2.1) may be re-expressed as

$$\hat{Y}_\theta = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{\theta,i} \tag{3.4}$$

with $\hat{Y}_{\theta,i} = \hat{Y}_{a,i}^A + \theta \hat{Y}_{ab,i}^A + (1-\theta) \hat{Y}_{ab,i}^B + \hat{Y}_{b,i}^B$ the Hartley estimator of sub-total Y_i over unit primary sampling unit u_i. We get $E(\hat{Y}_\theta \mid S_I) = \sum_{i \in S_I} d_{Ii} Y_i$, then

$$V(\hat{Y}_\theta) = V\left(\sum_{i \in S_I} d_{Ii} Y_i\right) + EV(\hat{Y}_\theta \mid S_I). \tag{3.5}$$

In (3.5), the first term of the right member does not depend on θ. Hartley's optimal estimator can, therefore, be calculated by minimizing the second term only. This gives:

$$\theta_{opt|S_I} = \frac{EV(\hat{Y}_{ab}^B \mid S_I) + ECov(\hat{Y}_{ab}^B, \hat{Y}_b^B \mid S_I) - ECov(\hat{Y}_a^A, \hat{Y}_{ab}^A \mid S_I)}{EV(\hat{Y}_{ab}^A \mid S_I) + EV(\hat{Y}_{ab}^B \mid S_I)}, \tag{3.6}$$

which can be estimated by

$$\hat{\theta}_{opt}=\frac{\hat{V}\left(\hat{Y}_{ab}^{B}\right)+\widehat{Cov}\left(\hat{Y}_{ab}^{B},\hat{Y}_{b}^{B}\right)-\widehat{Cov}\left(\hat{Y}_{a}^{A},\hat{Y}_{ab}^{A}\right)}{\hat{V}\left(\hat{Y}_{ab}^{A}\right)+\hat{V}\left(\hat{Y}_{ab}^{B}\right)} \tag{3.7}$$

by replacing each variance and covariance term with an unbiased estimator conditional on the first stage.

3.2 Kalton and Anderson estimator

With the sample design considered, we get $d_k^A = d_{Ii} d_{k|i}^A$ for any unit $k \in u_i^A$, and $d_k^B = d_{Ii} d_{k|i}^B$ for any unit $k \in u_i^B$. Therefore, the Kalton and Anderson estimator given in (2.4) can be re-expressed as

$$\hat{Y}_{KA} = \sum_{i \in S_I} d_{Ii} \hat{Y}_{KA,i} \tag{3.8}$$

with $\hat{Y}_{KA,i} = \sum_{k \in S^A} d_{k|i}^A m_{k|i}^A y_k + \sum_{k \in S^B} d_{k|i}^B m_{k|i}^B y_k$ the Kalton and Anderson estimator of the sub-total Y_i, where

$$m_{k|i}^A = \begin{cases} 1 & \text{if } k \in a \cap u_i, \\ \dfrac{d_{k|i}^B}{d_{k|i}^A + d_{k|i}^B} & \text{if } k \in ab \cap u_i, \end{cases} \quad \text{and} \quad m_{k|i}^B = \begin{cases} 1 & \text{if } k \in b \cap u_i, \\ \dfrac{d_{k|i}^A}{d_{k|i}^A + d_{k|i}^B} & \text{if } k \in ab \cap u_i. \end{cases}$$

3.3 Bankier estimator

With the sampling design considered, we get $\pi_k^{HT} = \pi_{Ii}\left(\pi_{k|i}^A + \pi_{k|i}^B - \pi_{k|i}^A \pi_{k|i}^B\right)$ for any $k \in u_i$. Therefore, the Bankier estimator given in (2.5) can be re-expressed as

$$\hat{Y}_{HT} = \sum_{i \in S_I} d_{Ii} \hat{Y}_{HT,i} \tag{3.9}$$

with $\hat{Y}_{HT,i} = \sum_{k \in S_i^A \cup S_i^B} \left(y_k / \pi_{k|i}^{HT}\right)$ the Bankier estimator for the sub-total Y_i, and $\pi_{k|i}^{HT} = \pi_{k|i}^A$ if $k \in a$, $\pi_{k|i}^{HT} = \pi_{k|i}^B$ if $k \in b$, $\pi_{k|i}^{HT} = \pi_{k|i}^A + \pi_{k|i}^B - \pi_{k|i}^A \pi_{k|i}^B$ if $k \in ab$.

Each of the three estimators examined is obtained by applying the estimation method PSU by PSU, conditional on the first stage. This result is particularly attractive for Hartley’s optimal method, since the optimal coefficient estimator given in (3.7) only requires variance estimators conditional on the first stage.

4 Simulation study

We are using artificial populations proposed by Saigo (2010). We generate two populations, each containing $M = 200$ primary sampling units grouped in $H = 4$ strata U_{Ih} of size $M_h = 50$. Each primary sampling unit u_{hi} contains $N_{hi} = 100$ secondary units. In each population, we generate for each primary sampling unit $u_{hi} \in U_{Ih}$:

$$\mu_{hi} = \mu_h + \sigma_h v_{hi} \tag{4.1}$$

where the values μ_h and σ_h are those used by Saigo (2010). The term σ_h^2 makes it possible to control dispersion between the primary sampling units. The v_{hi} are iid, generated according to a standard normal distribution $N(0,1).$ For each unit $k \in u_{hi},$ we then generate the value y_k according to the model

$$y_k = \mu_{hi} + \left\{\rho^{-1}(1-\rho)\right\}^{0.5} \sigma_h v_k, \tag{4.2}$$

where the v_k are iid, generated according to standard normal distribution. The variance term in the model (4.2) can give an intra-cluster correlation coefficient approximately equal to $\rho.$ In particular, the larger the ρ coefficient, the less the values y_k are dispersed in the primary sampling units. We use $\rho = 0.2$ for population 1 and $\rho = 0.5$ for population 2, which reflects less dispersion of the variable y in population 2. The sampling frame U_A corresponds to all secondary units, and the corresponding part of u_{hi} is $u_{hi}^A = u_{hi},$ of size $N_{hi}^A = N_{hi}.$ For each secondary unit $k,$ a value u_k is generated according to uniform distribution over $[0,1].$ The sampling frame U_B corresponds to the secondary units k such that $u_k \le 0.5,$ and the corresponding part of u_{hi} is $u_{hi}^B = u_{hi} \cap U_B$ of size $N_{hi}^B.$ This gives, therefore, the situation where $ab = U_B$ and $b = \varnothing.$ The framework selected in the simulations is the one used in the INSEE household surveys, with expansion to target a specific sub-population. For these surveys, a sample S_I of communes (or groups of communes) is first selected in the first stage. A sub-sample S_i^A of dwellings is then selected in each $u_i \in S_I;$ the pooled sample $S^A = \bigcup_{u_i \in S_I} S_i^A$ represents the entire population of dwellings $U_A = U.$ A second sub-sample S_i^B of dwellings is then selected from within a sub-population of each $u_i \in S_I,$ in order to target a specific sub-population U_B (for example, dwellings located in a Sensitive Urban Area); the pooled sample $S^B = \bigcup_{u_i \in S_I} S_i^B$ represents only the targeted sub-population $U_B.$

In each of the two populations created, several samplings are taken concurrently; Table 4.1 presents for each population the eight possible combinations of sample sizes per stratum in the first and second stage, as well as the values μ_h and $\sigma_h.$ In the first stage, we select independently in each stratum U_{Ih} : either a sample S_{Ih} of $m_h = 5$ primary sampling units by simple random sampling; or a sample S_{Ih} of $m_h = 25$ primary sampling units by simple random sampling. In the second stage, we select in each $u_{hi} \in S_{Ih}$: either a sample S_{hi}^A of size $n_{hi}^A = 10$ by simple random sampling in $u_{hi}^A;$ or a sample S_{hi}^A of size $n_{hi}^A = 40$ by simple random sampling in $u_{hi}^A.$ In the second stage, we also select in each $u_{hi} \in S_{Ih}$: either a sample S_{hi}^B of size $n_{hi}^B = 5$ by simple random sampling in $u_{hi}^B;$ or a sample S_{hi}^B of size $n_{hi}^B = 20$ by simple random sampling in $u_{hi}^B.$ Also we note $f_{hi}^A = \left(N_{hi}^A\right)^{-1} n_{hi}^A$ and $f_{hi}^B = \left(N_{hi}^B\right)^{-1} n_{hi}^B$ the sampling rates in u_{hi}^A and $u_{hi}^B.$

Table 4.1
Parameters used in each stratum to generate both populations and select samples

	Sample Sizes per Stratum			**Parameters**							
				Stratum 1		**Stratum 2**		**Stratum 3**		**Stratum 4**	
	m_h	n_{hi}^A	n_{hi}^B	μ_h	σ_h	μ_h	σ_h	μ_h	σ_h	μ_h	σ_h
Population 1	5 or 25	10 or 40	5 or 20	200	20	150	15	120	12	100	10
Population 2	5 or 25	10 or 40	5 or 20	200	10	150	7.5	120	6	100	5

For each sample, Hartley's estimator given in (3.4) is calculated with either $\theta = 1/2$ (HART1), or for value of θ the optimal coefficient estimator given in (3.7) (HART2), with

$$\hat{V}\left(\hat{Y}_{ab}^{A}\right) = \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{Ih}} \left(N_{hi}^{A}\right)^2 \frac{1-f_{hi}^{A}}{n_{hi}^{A}\left(n_{hi}^{A}-1\right)} \sum_{k\in S_{hi}^{A}} \left\{ y_k 1\left(k\in ab\right) - \bar{y}_{ab;S_{hi}^{A}} \right\}^2,$$

$$\hat{V}\left(\hat{Y}_{ab}^{B}\right) = \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{Ih}} \left(N_{hi}^{B}\right)^2 \frac{1-f_{hi}^{B}}{n_{hi}^{B}\left(n_{hi}^{B}-1\right)} \sum_{k\in S_{hi}^{B}} \left\{ y_k 1\left(k\in ab\right) - \bar{y}_{ab;S_{hi}^{B}} \right\}^2,$$

$$\widehat{Cov}\left(\hat{Y}_{a}^{A},\hat{Y}_{ab}^{A}\right) = \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{Ih}} \left(N_{hi}^{A}\right)^2 \frac{1-f_{hi}^{A}}{n_{hi}^{A}\left(n_{hi}^{A}-1\right)} \sum_{k\in S_{hi}^{A}} \left\{ y_k 1\left(k\in a\right) - \bar{y}_{a;S_{hi}^{A}} \right\}\left\{ y_k 1\left(k\in ab\right) - \bar{y}_{ab;S_{hi}^{A}} \right\},$$

noting $\bar{y}_{d;V}$ the average of variable $y_k 1\left(k\in d\right)$ on a subset $V\subset U.$ For each sample, the Kalton and Anderson estimator (KALT) given in (3.8) is also calculated, as well as the Bankier estimator (BANK) given in (3.9), and the Horvitz-Thompson estimator $\hat{Y}^{A}$ based on the single sample S^{A} (HTA). The sampling procedure is repeated 10,000 times. To measure the bias of an estimator $\hat{Y},$ we calculate its relative Monte Carlo bias

$$RB_{MC}\left(\hat{Y}\right) = \frac{E_{MC}\left(\hat{Y}\right)-Y}{Y}\times 100$$

with $E_{MC}\left(\hat{Y}\right) = \left(1/10,000\right)\sum_{b=1}^{10,000}\hat{Y}_{(b)},$ and $\hat{Y}_{(b)}$ the value of estimator $\hat{Y}$ for sample *b*. To measure the variability of $\hat{Y},$ we calculate its Monte Carlo mean square error

$$MSE_{MC}\left(\hat{Y}\right) = \frac{1}{10,000}\sum_{b=1}^{10,000}\left(\hat{Y}_{(b)}-Y\right)^2.$$

The results are given in Table 4.2. As emphasized by a referee, the performances of the HTA estimator do not depend on the sample size n_{hi}^{B} chosen. For consistency, Table 4.2 indicates the results obtained in the simulations with $n_{hi}^{B}=5$ only. For identical sample sizes m_h and identical $n_{hi}^{A},$ the same results are reported in the case n_{hi}^{B} = 20.

All estimators are virtually unbiased. The HART2 estimator gives better results in terms of mean squared error, as could be expected. The HTA estimator gives almost equivalent results. This result is explained by the fact that the optimal coefficient is near 1 (in the simulations, $\hat{\theta}_{opt}$ is between 0.80 and $1.06),$ and that in this case, the formula (2.1) shows that the HART2 and HTA estimators are very close: In the appendix we present some general conditions under which this property is approximately checked. Of the three estimators, HART1 yields the best results, with a mean square error lower than or equivalent to that of KALT and BANK in 11 out of 16 cases.

Table 4.2
Relative bias and mean squared error of five estimators

				HART1		HART2		KALT		BANK		HTA	
Pop.	m_h	n_{hi}^A	n_{hi}^B	**RB**	**MSE**	**RB**	**MSE**	**RB**	**MSE**	**RB**	**MSE**	**RB**	**MSE**
				(%)	$\times 10^9$	**(%)**	$\times 10^9$	**(%)**	$\times 10^9$	**(%)**	$\times 10^9$	**(%)**	$\times 10^9$
1	5	10	5	0.05	7.76	0.01	5.70	0.05	7.79	0.06	8.56	0.04	5.75
1	5	10	20	0.01	7.57	-0.05	5.57	0.03	11.36	0.04	12.75	0.04	5.75
1	5	40	5	0.01	5.01	-0.02	4.51	-0.02	4.57	-0.02	4.81	-0.02	4.52
1	5	40	20	0.00	4.65	-0.01	4.33	0.00	4.66	0.00	5.22	-0.02	4.52
1	25	10	5	-0.03	1.19	-0.02	0.78	-0.03	1.20	-0.02	1.34	-0.01	0.78
1	25	10	20	-0.01	1.17	0.00	0.78	-0.03	1.94	-0.03	2.22	-0.01	0.78
1	25	40	5	0.00	0.62	0.01	0.51	0.00	0.52	0.00	0.57	0.01	0.51
1	25	40	20	0.02	0.58	0.01	0.51	0.02	0.58	0.02	0.68	0.01	0.51
2	5	10	5	0.00	3.59	0.01	1.15	0.00	3.56	0.02	4.38	0.01	1.15
2	5	10	20	0.00	3.60	-0.02	1.15	0.00	7.38	0.00	8.76	0.01	1.15
2	5	40	5	0.00	1.48	0.01	1.07	0.00	1.13	0.01	1.35	0.01	1.07
2	5	40	20	0.00	1.49	-0.01	1.09	0.00	1.49	0.00	2.03	0.01	1.07
2	25	10	5	0.00	0.63	0.00	0.14	0.00	0.63	0.00	0.78	0.00	0.14
2	25	10	20	0.00	0.62	0.00	0.13	0.00	1.38	0.00	1.67	0.00	0.14
2	25	40	5	0.00	0.20	0.00	0.12	0.00	0.13	0.00	0.18	0.00	0.12
2	25	40	20	0.00	0.20	0.00	0.12	0.00	0.20	0.01	0.31	0.00	0.12

For each estimator, all other things being equal, the mean square error is lower in population 2 than in population 1. This result comes from the fact that the variance due to the first-stage selection, which is the same for each estimator and is

$$V\left(\sum_{i\in S_I} d_{Ii} Y_i\right) = \sum_{h=1}^{H} M_h^2 \left(\frac{1}{m_h} - \frac{1}{M_h}\right) S_{Y;U_{Ih}}^2, \tag{4.3}$$

is larger in population 1: the dispersion term $S_{Y;U_{Ih}}^2 = \left(M_h - 1\right)^{-1} \sum_{u_i \in U_{Ih}} \left(Y_i - \overline{Y}_{U_{Ih}}\right)^2$ increases with σ_h^2 and, to a lesser degree, increases when ρ decreases. The mean square error decreases for each estimator when the number m_h of primary sampling units selected in each stratum increases, since in this case the common variance term given in (4.3) decreases. Similarly, the mean square error decreases for each estimator when n^A increases, since in this case the variance due to the second stage of selection decreases. For the HART1 and HART2 estimators, the mean square error is stable when n^B increases, and more surprisingly for the KALT and BANK estimators the mean square error increases when n^B increases. This somewhat counterintuitive result is due to the convergence of two facts. On one hand, the contribution of sample S^B to the variance due to the second stage of selection is low: the increase of n^B may reduce this variance, but even in this case, overall reduction of the variance is marginal. On the other hand, with the KALT and BANK estimators, the contribution of sample S^A to the variance due to the second stage of selection increases when n^B increases.

In the case of KALT, the estimator can be re-expressed

$$\hat{Y}_{KA} = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{i \in S_{Ih}} \hat{Y}_{KA,i}$$

with

$$\hat{Y}_{KA,i} = \frac{1}{f_{hi}^{A}} \sum_{k \in S_i^A} m_{k|i}^{A} y_k + \frac{1}{f_{hi}^{A} + f_{hi}^{B}} \sum_{k \in S_i^B} y_k \quad \text{and} \quad m_{k|i}^{A} = \begin{cases} 1 & \text{if } k \in a \cap u_i, \\ \dfrac{f_{hi}^{A}}{f_{hi}^{A} + f_{hi}^{B}} & \text{if } k \in ab \cap u_i. \end{cases} \tag{4.4}$$

In (4.4), the dispersion of the variable $m_{k|i}^{A}$ (and therefore, that of $m_{k|i}^{A}\ y_k$) increases when the factor $f_{hi}^{A} \big/ \left(f_{hi}^{A} + f_{hi}^{B} \right)$ moves away from 1. This factor is near 1 when f_{hi}^{B} is small compared to f_{hi}^{A} (and therefore, if n^B is small compared to n^A), but moves away from 1 when n^B increases. Note that the variance (conditional on S_I) of the second term of $\hat{Y}_{KA,i}$ is equal to

$$V\left(\frac{1}{f_{hi}^{A} + f_{hi}^{B}} \sum_{k \in S_i^B} y_k \middle| S_I \right) = \left(N_{hi}^{A} \right)^2 N_{hi}^{B} \times \frac{n_{hi}^{B} \left(N_{hi}^{B} - n_{hi}^{B} \right)}{\left(N_{hi}^{B} n_{hi}^{A} + N_{hi}^{A} n_{hi}^{B} \right)^2} \times S_{u_{hi}^B}^2$$

with $S_{u_{hi}^B}^2 = \left(N_{hi}^{B} - 1 \right)^{-1} \sum_{k \in u_{hi}^B} \left(y_k - \bar{y}_{u_{hi}^B} \right)^2$. This variance does not necessarily decrease when n_{hi}^{B} increases. For example, one of the cases considered in the simulations corresponds to $N_{hi}^{A} = 100$, $N_{hi}^{B} \simeq 50$ and $n_{hi}^{A} = 40$. In this case, the term $n_{hi}^{B} \left(N_{hi}^{B} - n_{hi}^{B} \right) \big/ \left(N_{hi}^{B} n_{hi}^{A} + N_{hi}^{A} n_{hi}^{B} \right)^2$ attains its maximum value for $n_{hi}^{B} = 11$.

In the case of BANK, the estimator can be re-expressed

$$\hat{Y}_{HT} = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{i \in S_{Ih}} \hat{Y}_{HT,i}$$

with

$$\hat{Y}_{HT,i} = \sum_{k \in S_i^A \cup S_i^B} \frac{y_k}{\pi_{k|i}^{HT}} \quad \text{and} \quad \pi_{k|i}^{HT} = \begin{cases} f_{hi}^{A} & \text{if } k \in a, \\ f_{hi}^{A} + f_{hi}^{B} \left(1 - f_{hi}^{A} \right) & \text{if } k \in ab. \end{cases} \tag{4.5}$$

In (4.5), dispersion of the variable $\pi_{k|i}^{HT}$ increases when the factor $f_{hi}^{B} \left(1 - f_{hi}^{A} \right)$ increases. This factor is close to 0 when n_{hi}^{B} (and, therefore, f_{hi}^{B}) is low, but increases when n_{hi}^{B} increases.

5 Conclusion

We examined the Hartley (1962), Kalton and Anderson (1986) and Bankier (1986) estimators to pool the samples resulting from two survey waves. More particularly, we studied the case where the first sample represents the entire population (completely representative sample), while the second represents only a part (partially representative sample). Within the framework considered in the simulations (also see the Appendix for a more general framework), using the partially representative sample did not improve accuracy: if its size increases, the accuracy of the estimators in the Hartley class remains stable or improves slightly, while the accuracy of the Kalton and Anderson and Bankier estimators is worsened. Hartley’s optimal estimator itself, although more complex to calculate, offers accuracy that is only slightly improved as compared to the classic Horvitz-Thompson estimator calculated on the fully representative sample. Although our simulation study is limited, the results suggest that the estimator should be chosen carefully when there are multiple survey frames, and that a simple estimator is sometimes preferable, even if it uses only part of the information collected.

Acknowledgements

The authors would like to thank an associate editor and referee for their careful reading and comments, which helped to significantly improve the article, and David Haziza for the useful discussions.

Appendix

A1. Comparison of Hartley’s optimal estimator and the Horvitz-Thompson estimator

Let us take the framework and notations from Section 4: samples S^A and S^B are selected using a two-stage frame with common first stage selection. Stratified simple random sampling is used at the first stage, and simple random sampling in each primary sampling unit at the second stage. The sampling frame U_A corresponds to the entire population, while the sampling frame U_B covers only part of the population.

With Hartley’s optimal estimator, the formula (3.6) gives

$$\theta_{opt|S_I} = \frac{EV\left(\hat{Y}_{ab}^{B} \mid S_I\right) - ECov\left(\hat{Y}_{a}^{A}, \hat{Y}_{ab}^{A} \mid S_I\right)}{EV\left(\hat{Y}_{ab}^{B} \mid S_I\right) + EV\left(\hat{Y}_{ab}^{A} \mid S_I\right)}.$$

After some calculation, we get

$$EV\left(\hat{Y}_{ab}^{A} \mid S_I\right) = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{u_{hi}\in U_{Ih}} \left(N_{hi}\right)^2 \frac{1-f_{hi}^{A}}{n_{hi}^{A}} \left\{ \frac{N_{hi}^{B}-1}{N_{hi}-1} S_{u_{hi}^{B}}^{2} + \frac{N_{hi}^{B}\left(N_{hi}-N_{hi}^{B}\right)\left(\bar{y}_{u_{hi}^{B}}\right)^2}{N_{hi}\left(N_{hi}-1\right)} \right\}, \qquad (A.1)$$

$$-ECov\left(\hat{Y}_{a}^{A}, \hat{Y}_{ab}^{A} \mid S_I\right) = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{u_{hi}\in U_{Ih}} \left(N_{hi}\right)^2 \frac{1-f_{hi}^{A}}{n_{hi}^{A}} \left\{ \frac{N_{hi}^{B}\left(\bar{y}_{u_{hi}^{B}}\right)\left(N_{hi}\bar{y}_{u_{hi}} - N_{hi}^{B}\bar{y}_{u_{hi}^{B}}\right)}{N_{hi}\left(N_{hi}-1\right)} \right\}$$

with $\bar{y}_{u_{hi}} = \left(N_{hi}\right)^{-1} \sum_{k\in u_{hi}} y_k$, $\bar{y}_{u_{hi}^{B}} = \left(N_{hi}^{B}\right)^{-1} \sum_{k\in u_{hi}^{B}} y_k$ and $S_{u_{hi}^{B}}^{2} = \left(N_{hi}^{B}-1\right)^{-1} \sum_{k\in u_{hi}^{B}} \left(y_k - \bar{y}_{u_{hi}^{B}}\right)^2$.

The Horvitz-Thompson estimator based on the single sample S^A and Hartley’s optimal estimator agree if the coefficient $\theta_{opt|S_I}$ is equal to 1, which is the case if $EV\left(\hat{Y}_{ab}^{A} \mid S_I\right) = -ECov\left(\hat{Y}_{a}^{A}, \hat{Y}_{ab}^{A} \mid S_I\right)$. This condition will be verified in particular if in (A.1) the terms between the brackets agree for each primary sampling unit u_{hi}. We get therefore $\theta_{opt|S_I} \simeq 1$ if

$$\forall \quad u_{hi}\in U_I \quad \frac{N_{hi}\left(N_{hi}^{B}-1\right)}{N_{hi}^{B}} \frac{S_{u_{hi}^{B}}^{2}}{\bar{y}_{u_{hi}^{B}}\left(N_{hi}\bar{y}_{u_{hi}} - N_{hi}^{B}\bar{y}_{u_{hi}^{B}}\right)} + \frac{\left(N_{hi}-N_{hi}^{B}\right)\bar{y}_{u_{hi}^{B}}}{N_{hi}\bar{y}_{u_{hi}} - N_{hi}^{B}\bar{y}_{u_{hi}^{B}}} \simeq 1. \qquad (A.2)$$

Let us suppose that the mean value of y is approximately the same in the frames U_A and U_B for each primary sampling unit, i.e. that $\forall\, u_{hi}\in U_I \quad \bar{y}_{u_{hi}^{B}} \simeq \bar{y}_{u_{hi}}$. Then, the condition (A.2) will be verified approximately if $\forall\, u_{hi}\in U_I \quad cv_{u_{hi}^{B}}^{2}$ is close to 0, with $cv_{u_{hi}^{B}} = \sqrt{S_{u_{hi}^{B}}^{2}} \Big/ \bar{y}_{u_{hi}^{B}}$.

In summary, the Horvitz-Thompson estimator based on the single sample S^A and Hartley's optimal estimator will be close if within each primary sampling unit u_{hi}: (a) there is not much difference in the mean value of y between the two bases, and (b) the variable y has low dispersion within u_{hi}^B. In the simulations, the condition (a) is approximately met since the distribution of individuals between the sampling frames U_A and U_B is completely random; the condition (b) is approximately met with values of $cv_{u_{hi}^B}^2$ varying from 0.02 to 0.10 for population 1, and from 0.001 to 0.005 for population 2.

References

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, p.1074-1079.

Bourdalle, G., Christine, M. and Wilms, L. (2000). Échantillons maître et emploi. *Série INSEE Méthodes*, 21, p. 139-173.

Hansen, M.H. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, p. 333-362.

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, p. 203-206.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, p. 663-685.

Kalton, G. and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, A*, 149, p. 65-82.

Lavallée, P. (2002). *Le sondage indirect, ou la méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles (Belgium) and Éditions Ellipses (France).

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

Lohr, S.L. (2007). Recent developments in multiple frame surveys. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 3257-3264.

Lohr, S.L. (2009). Multiple frame surveys. In *Handbook of Statistics, Sample Surveys: Design, Methods and Applications*, Eds., D. Pfeffermann and C.R. Rao. Amsterdam: North Holland, Vol. 29A, p. 71-88.

Lohr, S.L. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology,* Vol.37 no.2, p. 197-213.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology*, Vol.33 no.2, p. 151-157.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics,* 3, p. 169-175.

Rao, J.N.K. and Wu, C. (2010). Pseudo-empirical likelihood inference for dual frame surveys. *Journal of the American Statistical Association*, 105, p. 1494-1503.

Saigo, H. (2010). Comparing four bootstrap methods for stratified three-stage sampling. *Journal of Official Statistics*, Vol. 26, No. 1, 2010, p. 193–207.