Journal of Official Statistics, Vol. 27, No. 4, 2011, pp. 651–668

On Variance Estimation for the French Master Sample

Guillaume Chauvet[1]

It is common practice to take advantage of auxiliary information collected from a large survey to improve the estimation for a smaller related survey. A comparable technique was used to draw the French Master Sample conditionally on the rotation groups of the New Census chiefly for practical reasons, that is, to benefit from a recent sampling frame. In this article, we are interested in the case where a sampling design, including the sampling units themselves and their selection probabilities, is defined conditionally on an external survey. More specifically, we focus on variance estimation for estimators arising from the French Master Sample.

Key words: Auxiliary information; balanced sampling; conditional sampling design; cube method; expansion estimator; primary sampling unit; rotation group.

1. Introduction

The Master Sample selected by the Institut National de la Statistique et des Etudes Economiques (INSEE) is a sample of dwellings used as a sampling frame for household surveys. Until 2004, the Master Sample was obtained, partly by sampling in the census of 1999, partly by using the New Dwellings Sampling Frame (NDSF) so as to represent the housing built since 1999. This approach involved following (at least) one part of the new dwellings, which resulted in a nonnegligible increase of the sampling costs.

Since 2004, the comprehensive Census of Population, conducted approximately every ten years, has been replaced by census surveys conducted annually. The detailed methodology is described in Godinot (2005). In each French region, small municipalities (less than 10,000 inhabitants in 1999) are randomly partitioned into five rotation groups by means of balanced sampling with equal probabilities. Each year, all the municipalities within one rotation group are surveyed, so that all the municipalities in the region are surveyed within a cycle of five years. Each large municipality (10,000 inhabitants or more in 1999) is the subject of an independent sampling design and is stratified according to the type of address (large addresses, new addresses, or other addresses). In each stratum the addresses are divided into five rotation groups. Each year, all the addresses within one rotation group (for the strata of large addresses and new addresses) or within a subsample (for the stratum of other addresses) are surveyed. After one cycle of five years, approximately 40% of the addresses in each large municipality are surveyed.

[1] ENSAI, Campus de Ker Lann, rue Blaise Pascal BP 37203, 35170 Bruz Cedex, France. Email: Guillaume.Chauvet@ensai.fr. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the INSEE.

Acknowledgments: The author is grateful to an Associate Editor and to two referees for their kind and detailed comments that led to a considerable improvement of the article.

The use of census surveys requires a change in the drawing of the Master Sample, since there no longer exists a sampling frame which gives, to date, the exact state of the housing. The selection of the Master Sample is itself subject to some constraints. First, household surveys carried out at year $t + 1$ must be selected in the census samples surveyed at year t. This provides a recent sampling frame, and avoids the extra cost of a specific system to cover the new dwellings. On the other hand, the principle of multistage sampling used for the previous Master Sample (Bourdalle et al. 2000) to reduce the survey costs is preserved, and a sample of Primary Sampling Units (PSUs) is selected. In the particular case of the Master Sample, these PSUs are denoted as Interviewer Action Areas (IAAs) see Berlemont et al. (2009), Christine and Faivre (2009). In each French region, these IAAs are built as follows. Each large municipality stands for one IAA, while small municipalities are aggregated to create an IAA. A sample of IAAs is then selected with probabilities proportional to size. To simplify, we will only focus on the case of IAAs emerging as aggregations of small municipalities in the remainder of the article, since the case of large municipalities does not involve any specific technical difficulties.

The joining of the two former constraints results in a further difficulty: a selected IAA must contain enough dwellings belonging to the rotation group surveyed at year t for the household surveys specifically conducted at year $t + 1$. For example, if 200 dwellings are needed in a particular IAA for all household surveys to be conducted in 2006, then the corresponding IAA must contain at least 200 dwellings located in the small municipalities from the rotation group surveyed in 2005. That is, the IAAs need to be defined conditionally on the rotation groups of the census. More specifically, the following constraint was imposed: each IAA must contain at least 300 dwellings in each of the five rotation groups. In summary, not only the Master Sample is selected conditionally on the new census, but the IAAs themselves as well as their probabilities of selection in the Master Sample are defined conditionally on the new census.

In this article we consider the problem of estimation with the new Master Sample. In Section 2, we briefly introduce balanced sampling by means of the cube method, and variance estimation in case of Horvitz-Thompson estimation. The notation for the Master Sample is given in Section 3, and the sampling design is described. A more detailed presentation may be found in Berlemont et al. (2009) and Christine and Faivre (2009). Two variance estimators for an expansion type estimator are proposed in Section 4, and compared in Section 5 through a set of simulations. Some concluding remarks are given in Section 6.

2. Balanced Sampling and Variance Estimation

Let U denote a finite population of size N. Let y be a variable of interest which takes the value y_k for unit k in U. We are interested in estimating the total $t_y = \sum_{k \in U} y_k$ of the variable y. Let S denote a random sample selected in U by means of a sampling design $p(\cdot)$. Let $\pi_k = \Pr(k \in S)$ denote the inclusion probability of unit k, and $\pi_{kl} = \Pr(k, l \in S)$ denote the second-order inclusion probability. Write $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k, \ldots, \pi_N)'$ for the vector of inclusion probabilities and $\boldsymbol{I}(S) = (I_1, \ldots, I_k, \ldots, I_N)'$ for the vector of sample membership indicators, where $I_k = 1$ if $k \in S$ and 0 otherwise.

The Horvitz-Thompson estimator

$$\hat{t}_{y\pi} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k = \sum_{k \in S} \frac{y_k}{\pi_k} \tag{1}$$

estimates without bias the finite population total t_y.

The sampling design is *balanced* on a q-vector $\mathbf{x}_k$ of auxiliary variables if the Horvitz-Thompson estimator $\hat{t}_{\mathbf{x}\pi} = \sum_{k \in U} \mathbf{x}_k I_k / \pi_k$ equals the real total $t_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$ of the (vector) $\mathbf{x}$-variable. In what follows, we assume that the sample S is selected by means of the cube method (Deville and Tillé 2004), which makes possible the selection of balanced samples if the balancing constraints

$$\hat{t}_{\mathbf{x}\pi} = t_{\mathbf{x}} \tag{2}$$

are satisfied, and approximately balanced samples if these are only closely satisfied. A rejective-type procedure may alternatively be used, see Fuller (2009). The choice of optimal inclusion probabilities for balanced sampling is discussed in Tillé and Favre (2005) and Chauvet et al. (2011).

The cube method proceeds in two phases: the flight phase, in which the balancing constraints are maintained exactly, and the landing phase, in which the balancing constraints are successively relaxed until the complete sample is obtained. Several implementations of the cube method have been proposed in the literature; see Tillé (2006) and Tillé (2010). In what follows, we consider the implementation described, for example, in Breidt and Chauvet (2011). Algorithm 1 given in Appendix A covers both the flight phase and the landing phase. It is currently used in practice, since it permits the selection of a balanced sample in a reasonable amount of time, even if the number of balancing variables is large. This implementation proceeds in steps $t = 0, 1, \ldots, T(S)$ from $\boldsymbol{\pi}_0(S) = \boldsymbol{\pi}$ to $\boldsymbol{\pi}_{T(S)}(S) = \boldsymbol{I}(S)$, the final sample. At each step, one or more coordinates of $\boldsymbol{\pi}_t(S)$ are randomly rounded to 0 or 1, and remain there forever. During the flight phase, the balancing equations remain exactly respected. When exact balance is no longer possible, the constraints are relaxed successively in the landing phase.

The sampling design is assumed to be of fixed size, which is obtained by including the vector $\boldsymbol{\pi}$ of inclusion probabilities in the balancing variables $\mathbf{x}$. The variance of the Horvitz-Thompson estimator is then given by the Yates-Grundy (1953) formula:

$$V(\hat{t}_{y\pi}) = -\frac{1}{2} \sum_{k,l \in U: k \neq l} \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \tag{3}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$, and may be unbiasedly estimated by

$$v(\hat{t}_{y\pi}) = -\frac{1}{2} \sum_{k,l \in S: k \neq l} \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \tag{4}$$

if all π_{kl} are strictly positive. Second-order inclusion probabilities are, however, usually difficult to compute for a general balanced sampling design.

A variance estimator may be obtained from (4) and from a simulation-based approximation of the design variance-covariance matrix,

$$\Delta = [\Delta_{kl}]_{k,l\in U}.$$

Though a direct simulation-based approximation is possible (see Fattorini 2006; Thompson and Wu 2008), an approximation that uses the martingale structure of the cube algorithm may be used. More specifically, Breidt and Chauvet (2011) demonstrate that the Δ matrix is unbiasedly estimated by

$$\Delta^{MD} = \frac{1}{C}\sum_{c=1}^{C}\sum_{t=1}^{T(S_c)} \lambda_{1t}^*(S_c)\lambda_{2t}^*(S_c)\boldsymbol{u}_t(S_c)\boldsymbol{u}_t'(S_c) \tag{5}$$

where $S_1, \ldots, S_c, \ldots, S_C$ are C independent replicates of the sample S, selected by Algorithm 1. The corresponding variance estimator for a given sample S is then obtained by plugging (5) into (4), which leads to

$$v_{MD}\left(\hat{t}_{y\pi}\right) = -\frac{1}{2}\sum_{k,l\in S:k\neq l}\frac{\Delta_{kl}^{MD}}{\pi_{kl}^{MD}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2 \tag{6}$$

where $\pi_{kl}^{MD} = \Delta_{kl}^{MD} + \pi_k\pi_l$. The numerical results in Breidt and Chauvet (2011) show a good performance of this estimator as compared to a naive, simulation-based variance estimator that does not use the martingale structure, and an essentially unbiased variance estimation.

Alternatively, Deville and Tillé (2005) propose a variance approximation if the balanced sampling is performed with maximum entropy. Under the assumptions that (i) the sampling design is exactly balanced, and (ii) the Horvitz-Thompson estimator is approximately normally distributed under Poisson sampling, they derive the variance approximation

$$V_{DT}(\hat{t}_{y\pi}) = \frac{N}{N-q}\sum_{k\in U}\pi_k(1-\pi_k)\left(\frac{y_k}{\pi_k}-\frac{y_k^*}{\pi_k}\right)^2 \tag{7}$$

where $y_k^* = \mathbf{x}_k'\beta_{\mathbf{xy}}$ is a weighted prediction of y_k obtained with the q balancing variables $\mathbf{x}_k$, and

$$\beta_{\mathbf{x}y} = \left(\sum_{l\in U}\pi_l(1-\pi_l)\frac{\mathbf{x}_l}{\pi_l}\frac{\mathbf{x}_l'}{\pi_l}\right)^{-1}\sum_{l\in U}\pi_l(1-\pi_l)\frac{\mathbf{x}_l}{\pi_l}\frac{y_l}{\pi_l} \tag{8}$$

Other, slightly different, variance approximations are proposed by Deville and Tillé. A variance estimator is obtained from (7) using a plug-in principle, by substituting each total $\sum_{k\in U}(\cdot)$ with its Horvitz-Thompson estimator $\sum_{k\in S}(\cdot)/\pi_k$. The resulting estimator is

$$v_{DT}(\hat{t}_{y\pi}) = \frac{n}{n-q}\sum_{k\in S}(1-\pi_k)\left(\frac{y_k}{\pi_k}-\frac{y_k^p}{\pi_k}\right)^2 \tag{9}$$

where $y_k^p = \mathbf{x}'_k \hat{\beta}_{\mathbf{x}y}$ and

$$\hat{\beta}_{\mathbf{x}y} = \left(\sum_{l \in S} (1 - \pi_l) \frac{\mathbf{x}_l}{\pi_l} \frac{\mathbf{x}'_l}{\pi_l} \right)^{-1} \sum_{l \in S} (1 - \pi_l) \frac{\mathbf{x}_l}{\pi_l} \frac{y_l}{\pi_l} \tag{10}$$

The variance estimator given in (9) essentially ignores sampling variation due to the landing phase, and may be severely negatively biased if the variance due to the landing phase is appreciable and/or if the balancing variables have a large explanatory power for the variable of interest; see Breidt and Chauvet (2011).

3. Notation and Estimation for the Master Sample

In the remainder of the article, let U denote the finite population of N small municipalities in one French region and let y be a variable of interest. First, the small municipalities are randomly split into $R = 5$ rotation groups by selecting R nonoverlapping samples with equal probabilities, balanced on a q-vector $\mathbf{x}_k$ of auxiliary variables, following the technique presented in Tillé and Favre (2004, Section 4). Let $\alpha_{kr} = 1/R$ denote the probability for municipality k to be selected in the rotation group G_r, $r = 1, \ldots, R$. Also, let $\alpha_{kl,r}$ denote the probability that municipalities k and l are selected jointly in the rotation group G_r. Since both the simple random sampling design and the stratified simple random sampling with proportional allocation may be seen as particular cases of balanced sampling with equal probabilities, the case when the rotation groups are selected by either of these two particular sampling designs is covered with our set-up.

The small municipalities are then grouped to obtain a population U_I of M PSUs u_i, $i = 1, \ldots, M$, which in the particular case of the French Master Sample and in the remainder of the article will be denoted as Interviewer Action Areas (IAAs). These IAAs are built by aggregating contiguous small municipalities, so that each IAA contains at least 300 dwellings in each rotation group. A sample S_I of IAAs is then selected with probabilities proportional to size, and we note π_{Ii} for the probability for IAA u_i to be selected in S_I, conditional on G_r, $r = 1, \ldots, R$. Also, note π_{Iij} for the (conditional) probability that IAAs u_i and u_j are selected jointly in the sample S_I. The sample S_I is selected by balanced sampling by means of the cube method, with balancing variables

$$\mathbf{z}_i^0 = \sum_{k \in u_i} \mathbf{z}_k^0 \tag{11}$$

and

$$\tilde{\mathbf{z}}_i^* = \left((\tilde{\mathbf{z}}_{i1}^*)', \ldots, (\tilde{\mathbf{z}}_{ir}^*)', \ldots, (\tilde{\mathbf{z}}_{iR}^{*\,\prime}) \right)' \tag{12}$$

with

$$\tilde{\mathbf{z}}_{ir}^* = \sum_{k \in u_i} \mathbf{z}_k^* 1(k \in G_r) \ \text{ for } \ r = 1, \ldots, R$$

where $\mathbf{z}_k^0$ and $\mathbf{z}_k^*$ denote two sets of auxiliary variables known at the design stage for any municipality k, and $1(\cdot)$ is the indicator function. The balancing variables $\mathbf{z}_i^0$ are used to achieve a global balancing of the sample S_I, and henceforth to obtain a variance reduction.

The variables $\mathbf{z}_k^0$ vary from one French region to another, since the more IAAs in a region, the more balancing variables may be added in the sampling design. The balancing variables $\tilde{\mathbf{z}}_i^*$ are used to achieve a balancing of the sample S_I on each rotation group, so as to benefit from a balanced sampling frame for each household survey. The $\mathbf{z}_k^*$ vector included the number of main dwellings and the global tax income of the small municipality k, and the $\mathbf{z}_k^0$ vector included the number of dwellings in peri-urban areas, rural areas and urban areas (see Berlemont et al. 2009). The global vector of balancing variables is given by $\tilde{\mathbf{z}}_i = ((\mathbf{z}_i^0)', (\tilde{\mathbf{z}}_i^*)')'$.

Finally, the small municipalities available for selection in the household surveys (at year $t+1$) are those in a specific rotation group G_r (of the census at time t) within the IAAs selected in S_I. This sample will be denoted as S_r, so that we have

$$S_r = \{k \in U;\ \ k \in u_i \in S_I \ \text{ and } \ k \in G_r\}. \tag{13}$$

Since the exact inclusion probabilities for the small municipalities k to be included in the final sample S_r are unknown, we propose using an expansion estimator instead of the Horvitz-Thompson estimator (see Särndal et al. 1992, p. 347). Note that the total t_y may alternatively be written as

$$t_y = \sum_{u_i \in U_I} Y_i$$

where $Y_i = \sum_{k \in u_i} y_k$ denotes the total of the y-variable in the IAA u_i. The proposed expansion estimator is given by

$$\hat{t}_{yr} = \sum_{k \in S_r} \frac{y_k}{p_{kr}} \tag{14}$$

where $p_{kr} = \pi_{Ii}\ \ \alpha_{kr}$ for any unit k in u_i. This estimator may also be written as

$$\hat{t}_{yr} = \sum_{u_i \in S_I} \frac{\tilde{Y}_{ir}}{\pi_{Ii}} \tag{15}$$

where

$$\tilde{Y}_{ir} = \sum_{k \in u_i} \frac{y_k 1(k \in G_r)}{\alpha_{kr}}$$

denotes a weighted total of the y-variable for the small municipalities in u_i which also belong to the rotation group G_r. It follows from (15) that

$$E(\hat{t}_{yr} | G_1, \ldots, G_R) = \sum_{u_i \in U_I} \tilde{Y}_{ir} = \sum_{k \in G_r} \frac{y_k}{\alpha_{kr}} \equiv \tilde{t}_{yr} \tag{16}$$

where $E(\cdot | G_1, \ldots, G_R)$ stands for the expectation conditional on the rotation groups of the census. Then, since

$$\sum_{u_i \in U_I} \tilde{Y}_{ir} = \sum_{u_i \in U_I} \sum_{k \in u_i} \frac{y_k 1(k \in G_r)}{\alpha_{kr}} = \sum_{k \in U} \frac{y_k 1(k \in G_r)}{\alpha_{kr}} \tag{17}$$

we obtain

$$E(\hat{t}_{yr}) = EE(\hat{t}_{yr}|G_1, \ldots, G_R) = E\left(\sum_{k\in U}\frac{y_k 1(k \in G_r)}{\alpha_{kr}}\right) = t_y$$

so that $\hat{t}_{yr}$ is an unbiased estimator of the total t_y.

4. Variance Estimation for the Expansion Estimator

The variance of the expansion estimator is given by

$$V(\hat{t}_{yr}) = VE(\hat{t}_{yr}|G_1, \ldots, G_R) + EV(\hat{t}_{yr}|G_1, \ldots, G_R) = V_{NC} + V_{MS}$$

where $V(\cdot|G_1, \ldots, G_R)$ denotes the variance conditional on the rotation groups of the census. From Equation (16), we get

$$V_{NC} = V\left(\sum_{k\in G_r}\frac{y_k}{\alpha_{kr}}\right) = -\frac{1}{2}\sum_{k,l\in U:k\neq l}\Delta_{kl,r}\left(\frac{y_k}{\alpha_{kr}} - \frac{y_l}{\alpha_{lr}}\right)^2$$

where $\Delta_{kl,r} = \alpha_{kl,r} - \alpha_{kr}\alpha_{lr}$. That is, V_{NC} corresponds to the variance due to the random selection of the rotation group G_r of the new census. The term V_{MS} corresponds to the variance due to the random selection of the sample S_I of IAAs. Using Equation (15), we can write

$$V_{MS} = E\left[V\left(\sum_{u_i\in S_I}\frac{\tilde{Y}_{ir}}{\pi_{Ii}}\middle|G_1, \ldots, G_R\right)\right] = E\left[-\frac{1}{2}\sum_{u_i,u_j\in U_I:u_i\neq u_j}\Delta_{Iij}\left(\frac{\tilde{Y}_{ir}}{\pi_{Ii}} - \frac{\tilde{Y}_{jr}}{\pi_{Ij}}\right)^2\right]$$

where $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$. In the sequel, the two terms V_{NC} and V_{MS} are estimated separately.

A direct estimator for V_{NC} is

$$v_{NC}(\hat{t}_{yr}) = -\frac{1}{2}\sum_{k,l\in S_r:k\neq l}\frac{\Delta_{kl,r}}{\alpha_{kl,r}\alpha_{kl|G}}\left(\frac{y_k}{\alpha_{kr}} - \frac{y_l}{\alpha_{lr}}\right)^2 \quad (18)$$

where

$$\alpha_{kl|G} \equiv \Pr(k, l \in S_r|G_1, \ldots, G_R) = \pi_{Iij} \text{ if } k \in u_i,\ l \in u_j. \quad (19)$$

It is shown in Appendix B that the estimator in (18) is unbiased for V_{NC}, provided that all $\alpha_{kl,r}$ and $\alpha_{kl|G}$ are strictly positive. We now turn to the second term V_{MS}. From the definition

$$V_{MS} = EV(\hat{t}_{yr}|G_1, \ldots, G_R)$$

we see that it is sufficient to find an estimator for $V(\hat{t}_{yr}|G_1, \ldots, G_R)$, which we can write as

$$V\left(\sum_{u_i \in S_I} \frac{\tilde{Y}_{ir}}{\pi_{Ii}}|G_1, \ldots, G_R\right).$$

Consequently, a direct estimator for V_{MS} is

$$v_{MS}(\hat{t}_{yr}) = -\frac{1}{2}\sum_{u_i,u_j \in S_I : u_i \neq u_j} \frac{\Delta_{Iij}}{\pi_{Iij}}\left(\frac{\tilde{Y}_{ir}}{\pi_{Ii}} - \frac{\tilde{Y}_{jr}}{\pi_{Ij}}\right)^2 \tag{20}$$

and $v_{MS}(\hat{t}_{yr})$ estimates V_{MS} unbiasedly if all π_{Iij} are strictly positive. Unfortunately, estimators (18) and (20) require the knowledge of second-order inclusion probabilities which are usually difficult to compute. In Section 4.1, a simulation-based variance estimator is proposed. In Section 4.2, we follow Deville and Tillé (2005) in proposing an alternative variance estimator.

4.1. Simulation-based Variance Estimation

A variance estimator may be obtained from (18) and (20), using a simulation-based approximation of the design variance-covariance matrices

$$\Delta_r \equiv [\Delta_{kl,r}]_{k,l \in U}$$

and

$$\Delta_I \equiv [\Delta_{Iij}]_{u_i,u_j \in U_I}.$$

More specifically, let

$$\Delta_r^{MD} = \frac{1}{C}\sum_{c=1}^{C}\sum_{t=1}^{T(G_{rc})} \lambda_{1t}^*(G_{rc})\lambda_{2t}^*(G_{rc})\boldsymbol{u}_t(G_{rc})\boldsymbol{u}_t'(G_{rc}) \tag{21}$$

where $G_{r1}, \ldots, G_{rc}, \ldots, G_{rC}$ are C independent replicates of the rotation group G_r, selected by Algorithm 1. Similarly, let

$$\Delta_I^{MD} = \frac{1}{C}\sum_{c=1}^{C}\sum_{t=1}^{T(S_{Ic})} \lambda_{1t}^*(S_{Ic})\lambda_{2t}^*(S_{Ic})\boldsymbol{u}_t(S_{Ic})\boldsymbol{u}_t'(S_{Ic}) \tag{22}$$

where $S_{I1}, \ldots, S_{Ic}, \ldots, S_{IC}$ are C independent replicates of the sample S_I, selected by Algorithm 1 conditionally on the rotation groups $G_r, \ r = 1, \ldots, R$.

The first proposed variance estimator is

$$v_{MD}(\hat{t}_{yr}) = v_{MD,NC}(\hat{t}_{yr}) + v_{MD,MS}(\hat{t}_{yr}). \tag{23}$$

The term

$$v_{MD,MS}(\hat{t}_{yr}) = -\frac{1}{2}\sum_{u_i,u_j \in S_I : u_i \neq u_j} \frac{\Delta_{Iij}^{MD}}{\pi_{Iij}^{MD}}\left(\frac{\tilde{Y}_{ir}}{\pi_{Ii}} - \frac{\tilde{Y}_{jr}}{\pi_{Ij}}\right)^2$$

is obtained from (20) by replacing Δ_{Iij} with the corresponding entry Δ_{Iij}^{MD} from (22) and π_{Iij}

with $\pi_{Iij}^{MD} = \Delta_{Iij}^{MD} + \pi_{Ii}\ \ \pi_{Ij}$. The term

$$v_{MD,NC}(\hat{t}_{yr}) = -\frac{1}{2} \sum_{k,l \in S_r : k \neq l} \frac{\Delta_{kl,r}^{MD}}{\alpha_{kl,r}^{MD} \alpha_{kl|G}^{MD}} \left(\frac{y_k}{\alpha_{kr}} - \frac{y_l}{\alpha_{lr}} \right)^2$$

is obtained from (18) by replacing $\Delta_{kl,r}$ with the corresponding entry $\Delta_{kl,r}^{MD}$ from (21), $\alpha_{kl,r}$ with $\alpha_{kl,r}^{MD} = \Delta_{kl,r}^{MD} + \alpha_{kr}\ \ \alpha_{lr}$, and $\alpha_{kl|G}$ with $\alpha_{kl|G}^{MD} = \pi_{Iij}^{MD}$ for $k \in u_i$, $l \in u_j$.

4.2. *Maximum Entropy Variance Estimation*

We now propose an alternative, non computational-intensive variance estimator, following the approach suggested by Deville and Tillé (2005). We focus on the first term V_{NC} first. Since the rotation group G_r is selected by means of balanced sampling with inclusion probabilities α_{kr}, the variance estimator given by (9) leads us to think of

$$\tilde{v}_{DT,NC}(\hat{t}_{yr}) = \frac{n(G_r)}{n(G_r) - q} \sum_{k \in G_r} (1 - \alpha_{kr}) \left(\frac{y_k}{\alpha_{kr}} - \frac{y_k^p}{\alpha_{kr}} \right)^2 \tag{24}$$

as an estimator of V_{NC}. In this expression, $n(G_r)$ is the number of municipalities selected in the rotation group G_r, $y_k^p = \mathbf{x}'_k \hat{\beta}_{\mathbf{x}y}$ is a weighted prediction of y_k obtained with the q balancing variables $\mathbf{x}_k$, and

$$\hat{\beta}_{\mathbf{x}y} = \left(\sum_{l \in G_r} (1 - \alpha_{lr}) \frac{\mathbf{x}_l}{\alpha_{lr}} \frac{\mathbf{x}'_l}{\alpha_{lr}} \right)^{-1} \sum_{l \in G_r} (1 - \alpha_{lr}) \frac{\mathbf{x}_l}{\alpha_{lr}} \frac{y_l}{\alpha_{lr}}$$

Unfortunately, the variance estimator given in (24) can not be computed since the variable of interest y is known on the sample S_r only. A tractable variance estimator is obtained from (24) using a plug-in principle, by substituting each total $\sum_{k \in G_r} (\cdot)$ with its expansion estimator $\sum_{k \in S_r} (\cdot)/\pi_{Ii} = \sum_{k \in S_r} \alpha_{kr}(\cdot)/p_{kr}$. We obtain

$$v_{DT,NC}(\hat{t}_{yr}) = \frac{n(G_r)}{n(G_r) - q} \sum_{k \in S_r} \frac{\alpha_{kr}(1 - \alpha_{kr})}{p_{kr}} \left(\frac{y_k}{\alpha_{kr}} - \frac{\hat{y}_k^p}{\alpha_{kr}} \right)^2 \tag{25}$$

where $\hat{y}_k^p = \mathbf{x}'_k \hat{\hat{\beta}}_{\mathbf{x}y}$ and

$$\hat{\hat{\beta}}_{\mathbf{x}y} = \left(\sum_{l \in S_r} \frac{\alpha_{lr}(1 - \alpha_{lr})}{p_{lr}} \frac{\mathbf{x}_l}{\alpha_{lr}} \frac{\mathbf{x}'_l}{\alpha_{lr}} \right)^{-1} \sum_{l \in S_r} \frac{\alpha_{lr}(1 - \alpha_{lr})}{p_{lr}} \frac{\mathbf{x}_l}{\alpha_{lr}} \frac{y_l}{\alpha_{lr}}$$

After some algebra, the variance estimator given in (25) may be alternatively written as

$$v_{DT,NC}(\hat{t}_{yr}) = \hat{\tilde{v}}_{DT,NC}(\hat{t}_{yr}) - \frac{n(G_r)}{n(G_r) - q} \left(\hat{\hat{\beta}}_{\mathbf{x}y} - \hat{\beta}_{\mathbf{x}y} \right)' \hat{A}_r \left(\hat{\hat{\beta}}_{\mathbf{x}y} - \hat{\beta}_{\mathbf{x}y} \right) \tag{26}$$

where

$$\hat{\tilde{v}}_{DT,NC}(\hat{t}_{yr}) = \frac{n(G_r)}{n(G_r) - q} \sum_{k \in S_r} \frac{\alpha_{kr}(1 - \alpha_{kr})}{p_{kr}} \left(\frac{y_k}{\alpha_{kr}} - \frac{y_k^p}{\alpha_{kr}} \right)^2$$

and

$$\hat{A}_r = \sum_{k \in S_r} \frac{\alpha_{kr}(1-\alpha_{kr})}{p_{kr}} \frac{\mathbf{x}_k}{\alpha_{kr}} \frac{\mathbf{x}'_k}{\alpha_{kr}}$$

We have $E\big(\hat{\tilde{v}}_{DT,NC}(\hat{t}_{yr})|G_1, \ldots, G_R\big) = \tilde{v}_{DT,NC}(\hat{t}_{yr})$, and under mild regularity conditions, the second term in the right-hand side of (26) is of smaller order of magnitude than $\hat{\tilde{v}}_{DT,NC}(\hat{t}_{yr})$. Consequently, $v_{DT,NC}(\hat{t}_{yr})$ is an approximately (conditionally) unbiased estimator for $\tilde{v}_{DT,NC}(\hat{t}_{yr})$.

We now turn to the second term V_{MS}. Once again, from the definition

$$V_{MS} = EV(\hat{t}_{yr}|G_1, \ldots, G_R)$$

we see that it is sufficient to find an estimator for $V(\hat{t}_{yr}|G_1, \ldots, G_R)$, which we can write as

$$V\left(\sum_{u_i \in S_I} \frac{\tilde{Y}_{ir}}{\pi_{Ii}} \middle| G_1, \ldots, G_R \right),$$

using Equation (15). Since the sample S_I is also obtained by balanced sampling by means of the Cube method, the variance estimator of Deville and Tillé (2005) may still be used. That is, the variance due to the selection of the Master Sample is estimated by

$$v_{DT,MS}(\hat{t}_{yr}) = \frac{m(S_I)}{m(S_I) - q_1} \sum_{u_i \in S_I} (1 - \pi_{Ii}) \left(\frac{\tilde{Y}_{ir}}{\pi_{Ii}} - \frac{\tilde{Y}^p_{ir}}{\pi_{Ii}} \right)^2 \tag{27}$$

where $m(S_I)$ gives the number of IAAs selected in S_I, and $\tilde{Y}^p_{ir} = \tilde{\mathbf{z}}'_i \hat{\gamma}_{\mathbf{z}Y}$ is a weighted prediction of $\tilde{Y}_{ir}$ obtained with the q_I balancing variables $\tilde{\mathbf{z}}_i$, where

$$\hat{\gamma}_{\mathbf{z}Y} = \left(\sum_{u_j \in S_I} (1 - \pi_{Ij}) \frac{\tilde{\mathbf{z}}_j}{\pi_{Ij}} \frac{\tilde{\mathbf{z}}'_j}{\pi_{Ij}} \right)^{-1} \sum_{u_j \in S_I} (1 - \pi_{Ij}) \frac{\tilde{\mathbf{z}}_j}{\pi_{Ij}} \frac{\tilde{Y}_{jr}}{\pi_{Ij}}$$

The second proposed variance estimator is thus given by

$$v_{DT}(\hat{t}_{yr}) = v_{DT,NC}(\hat{t}_{yr}) + v_{DT,MS}(\hat{t}_{yr}). \tag{28}$$

5. A Simulation Study

We performed a limited simulation study to assess the performance of the proposed variance estimators. We used a population of $N = 1,235$ small municipalities in the French region of Brittany. The variables of interest, available from the Census of 1999, are given in Table 1. The objective was to estimate the variance of the expansion estimator $\hat{t}_{yr}$ of the totals of the y-variables of interest. The simulation was performed with the SAS software. Balanced sampling can be implemented using existing software for the cube method, such as the R function samplecube in the sampling library (Tillé and Matei 2008; R Development Core Team 2008), or the SAS Macro fastcube (Chauvet and Tillé 2005).

Table 1. Variables of interest in the simulation study

Variable	Description
POP22	Number of people in the department of Côtes d'Armor
POP29	Number of people in the department of Finistère
POP35	Number of people in the department of Ille et Vilaine
POP56	Number of people in the department of Morbihan
FOREIGN	Number of foreigners
FORM	Number of men, foreigners
FORWACT	Number of women in active, foreigners
FORMACT	Number of men in active, foreigners
SALMACT	Number of men in active, salaried
EMPMACT	Number of men in active, employed
SALACT	Number of people in active, salaried
EMPACT	Number of people in active, employed
NONDIP	Number of people, unqualified
NSALACT	Number of people in active, nonsalaried
NSALMACT	Number of men in active, nonsalaried
UNEMP	Number of people, unemployed

The sampling design used in the simulation was meant to mimic closely the exact sampling design used for the selection of the French Master Sample. The sampling process described in section 3 was repeated $B = 1,000$ times to obtain $R = 5$ rotation groups $G_{1,b}, \ldots, G_{R,b}$, a population $U_{I,b}$ of IAAs in which a sample $S_{I,b}$ was selected with inclusion probabilities proportional to the number of main dwellings, and the final sample $S_{r,b}$, for $b = 1, \ldots, B$. The balancing variables $\mathbf{x}_k$ used in the selection of the rotation groups included the inclusion probability, the number of households, the number of households in collective addresses, the number of people by gender, and the number of people in five age classes. In the selection of the sample $S_{I,b}$, the vector $\mathbf{z}_i^0$ was limited to the inclusion probability, and the vector $\tilde{\mathbf{z}}_i^*$ consisted of the number of main dwellings and the global tax income by rotation group.

In each sample $S_{r,b}$, we computed the expansion estimator $\hat{t}_{yr,b}$, the simulation-based variance estimators $v_{MD,NC}(\hat{t}_{yr,b})$ and $v_{MD,MS}(\hat{t}_{yr,b})$, the maximum entropy variance estimators $v_{DT,NC}(\hat{t}_{yr,b})$ and $v_{DT,MS}(\hat{t}_{yr,b})$. The Δ_r^{MD} matrix in (21) was obtained from a single, separate simulation of $C = 10,000$ samples. For any $b = 1, \ldots, B$, a $\Delta_{I,b}^{MD}$ matrix was obtained from (22) and from a simulation run of $C = 10,000$ independent replicates $S_{I,b1}, \ldots, S_{I,bC}$ of the sample $S_{I,b}$, selected by Algorithm 1 conditionally on the rotation groups $G_{r,b}$, $r = 1, \ldots, R$. For simplicity, we present the simulation results in the case $r = 1$ only.

As a measure of bias of a point estimator $\hat{\theta}$ of a parameter θ, we used the Monte Carlo percent relative bias (RB) given by

$$RB_{MC}(\hat{\theta}) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{\theta}_{(b)} - \theta}{\theta}$$

where $\hat{\theta}_{(b)}$ gives the value of the estimator for the bth sample. When $\theta = V_{NC}(\hat{t}_{yr})$, we have $\hat{\theta}$ equal to either $v_{DT,NC}(\hat{t}_{yr})$ or $v_{MD,NC}(\hat{t}_{yr})$. When $\theta = V_{MS}(\hat{t}_{yr})$, we have $\hat{\theta}$ equal to either

$v_{DT,MS}(\hat{t}_{yr})$ or $v_{MD,MS}(\hat{t}_{yr})$. The exact variances $V_{NC}(\hat{t}_{yr})$ and $V_{MS}(\hat{t}_{yr})$ were replaced by a Monte Carlo approximation, obtained through an independent run of $30,000 \times 50$ simulations. More precisely, we repeated $D = 30,000$ times the creation of $R = 5$ rotation groups $G_{1,d}, \ldots, G_{R,d}$, and of a population $U_{I,d}$ of IAAs, for $d = 1, \ldots, D$. For any d, $E = 50$ samples $S_{I,de}$ were selected and the final sample $S_{r,de}$ was obtained, for $e = 1, \ldots, E$. The Monte Carlo approximation of $V_{NC}(\hat{t}_{yr})$ is

$$V_{NC}^{MC}(\hat{t}_{yr}) = \frac{1}{D-1}\sum_{d=1}^{D}\left(\tilde{t}_{yr,d} - \bar{\tilde{t}}_{yr}\right)^2$$

where

$$\tilde{t}_{yr,d} = \sum_{k \in G_{r,d}} \frac{y_k}{\alpha_{kr}} \text{ and } \bar{\tilde{t}}_{yr} = \frac{1}{D}\sum_{d=1}^{D}\tilde{t}_{yr,d}.$$

The Monte Carlo approximation of $V_{MS}(\hat{t}_{yr})$ is

$$V_{MS}^{MC}(\hat{t}_{yr}) = \frac{1}{D}\sum_{d=1}^{D}\left[\frac{1}{E-1}\sum_{e=1}^{E}\left(\hat{t}_{yr,de} - \bar{\hat{t}}_{yr,d}\right)^2\right]$$

where

$$\bar{\hat{t}}_{yr,d} = \frac{1}{E}\sum_{e=1}^{E}\hat{t}_{yr,de}$$

for any $d = 1, \ldots, D$. The results are presented in Table 2 for the variance due to the selection of the Master Sample, along with the mean coefficient of determination (R^2) obtained by predicting $\tilde{Y}_{ir}/\pi_{Ii}$ with the balancing variables $\tilde{\mathbf{z}}_i$. The results are presented in Table 3 for the variance due to the selection of the rotation groups of the new census, along with the coefficient of determination (R^2) obtained by predicting y_k/α_{kr} with the balancing variables $\tilde{\mathbf{x}}_k$.

For the two terms of variance, the MD estimator systematically outperforms the DT estimator in terms of relative bias. The MD estimator is essentially unbiased in any case, with a relative bias lower than 5% for V_{MS}, and lower than 6% for V_{NC}. We note that the DT estimator is systematically negatively biased, which is consistent with the results in Breidt and Chauvet (2011) in indicating that the DT estimator fails to track the variance due to the landing phase. In case of V_{MS}, the relative bias of the DT estimator increases as the R^2 of the model increases, as the balancing variables have more and more explanatory power and as the relative importance of the variance due to the flight phase in the overall variance decreases. In the case of V_{NC}, the relative bias of the DT estimator is large, irrespective of the explanatory power of the balancing variables. Our interpretation is as follows. The landing phase is applied to a subgroup of units in the population only. This subgroup is obtained randomly (these are the units which remain after the flight phase), and even the number of units in this subgroup is random. Moreover, in case of the new census, the units in this subgroup may vary considerably in size, since the small municipalities have very different sizes. For example, the coefficient of variation for the municipality sizes in Brittany is equal to 102%. In case of the new census, there is

Table 2. *Relative bias for two estimators of the variance due to the selection of the Master Sample*

	POP22	POP29	POP35	POP56	FOREIGN	FORM	FORWACT	FORMACT
	% Rel. Bias RB_{MC}							
DT	−7.6	−9.0	−12.8	−7.9	−14.3	−13.5	−13.7	−16.5
MD	0.6	−0.5	0.8	2.0	−2.4	−0.7	−3.0	−2.9
	Coeff. of determination R^2							
	0.19	0.17	0.23	0.16	0.48	0.49	0.48	0.48
	SALMACT	EMPMACT	SALACT	EMPACT	NONDIP	NSALACT	NSALMACT	UNEMP
	% Rel. Bias RB_{MC}							
DT	−20.2	−22.7	−21.9	−24.4	−27.5	−31.9	−32.7	−42.7
MD	−4.4	−4.6	−3.7	−4.1	−1.6	−0.6	−0.4	−0.8
	Coeff. of determination R^2							
	0.61	0.68	0.66	0.72	0.83	0.86	0.87	0.91

Table 3. Relative bias for two estimators of the variance due to the selection of the new census

	POP22	POP29	POP35	POP56	FOREIGN	FORM	FORWACT	FORMACT
	% Rel. Bias RB_{MC}							
DT	−22.9	−18.3	−20.6	−19.7	−19.2	−21.9	−13.3	−23.9
MD	0.2	−0.9	0.3	−1.7	2.6	4.6	−1.8	3.3
	Coeff. of determination R^2							
	0.18	0.27	0.37	0.22	0.48	0.49	0.48	0.54
	SALMACT	EMPMACT	SALACT	EMPACT	NONDIP	NSALACT	NSALMACT	UNEMP
	% Rel. Bias RB_{MC}							
DT	−31.3	−31.1	−30.0	−30.2	−17.8	−15.9	−16.0	−36.0
MD	−6.0	−6.0	−5.7	−5.7	−1.3	0.2	0.6	0.0
	Coeff. of determination R^2							
	0.67	0.72	0.74	0.77	0.85	0.89	0.89	0.93

consequently a nonnegligible variance due to the landing phase for any variable of interest. In the case of the selection of the IAAs, the use of inclusion probabilities proportional to size makes it possible to account for a variation in size of the IAAs, and so to limit the variance due to the landing phase.

6. Conclusion

In this article, expansion estimation for the French Master Sample has been considered, and two variance estimators have been compared. The proposed simulation-based variance estimator is shown to perform well in terms of relative bias, while the maximum entropy variance estimator may be severely biased. Also, we noted that the variance due to the landing phase may be appreciable even if the balancing variables have a small explanatory power for the variables of interest.

To simplify the presentation, we focused on the heart of the new sampling design only, which involved a random selection of PSUs themselves defined conditionally on the rotation groups of the new census. Some further specific points must be taken into account to fully account for the actual sampling design of household surveys. First, the final sample of dwellings for a specific household survey is obtained via an additional stage of sampling inside the Master Sample. The global variance is thus estimated by summing one of the variance estimators proposed in Section 4, which accounts for the variance due to the selection of the Master Sample, and a variance estimator which accounts for the subsampling of dwellings. Secondly, the variance of the expansion estimator is reduced by calibrating on known totals. The effect of this calibration may simply be taken into account by a residual technique, that is, by replacing in the variance estimator the variable of interest y with the residual of the regression of the variable y on the vector of calibration variables (see Deville and Särndal 1992).

Appendix

A. Algorithm 1 for Implementation of the Cube Method

Define the balancing matrix $A = (\mathbf{a}_1, \ldots, \mathbf{a}_k, \ldots, \mathbf{a}_N)$, where $\mathbf{a}_k = \mathbf{x}_k/\pi_k$. First initialize with $\boldsymbol{\pi}_0(S) = \boldsymbol{\pi}$ and $A(0) = A$. Next, at time $t = 0, \ldots, T(S)$, repeat the three following steps.

Step 1:

Let $E(t) = F(t) \cap \mathrm{Ker}A(t)$, where

$$F(t) = \{\mathbf{v} \in \mathbb{R}^N : v_k = 0 \text{ if } \pi_{kt}(S) \text{ is an integer}\},$$

with $\boldsymbol{\pi}_t(S) = (\pi_{1t}(S), \ldots, \pi_{kt}(S), \ldots, \pi_{Nt}(S))'$. Then:

- If $E(t) \neq \{0\}$, generate any vector $\mathbf{u}_t(S) \neq \mathbf{0}$ in $E(t)$, random or not. Put $A(t+1) = A(t)$.
- If $E(t) = \{0\}$, let m_t denote the largest integer such that $F(t) \cap \mathrm{Ker}A_{m_t}(t) \neq \{0\}$, where $A_{m_t}(t)$ denotes the matrix given by the first m_t rows of $A(t)$. Generate any vector $\mathbf{u}_t(S) \neq 0$ in $F(t) \cap \mathrm{Ker}A_{m_t}(t)$, random or not. Put $A(t+1) = A_{m_t}(t)$.

Step 2:

Compute the scalars $\lambda_{1t}^*(S)$ and $\lambda_{2t}^*(S)$, which are the largest values of λ_{1t} and λ_{2t} such that

$$0 \le \boldsymbol{\pi}_t(S) + \lambda_{1t}\mathbf{u}_t(S) \le 1, 0 \le \boldsymbol{\pi}_t(S) - \lambda_{2t}\mathbf{u}_t(S) \le 1,$$

where the inequalities are interpreted element-wise. Note that $\lambda_{1t}^*(S) > 0$ and $\lambda_{2t}^*(S) > 0$.

Step 3:

Select $\boldsymbol{\pi}_{t+1}(S) = \boldsymbol{\pi}_t(S) + \boldsymbol{\delta}_t\,(S)$, where

$$\boldsymbol{\delta}_t(S) = \begin{cases} \lambda_{1t}^*(S)\mathbf{u}_t(S) & \text{with probability} \quad q(t) \\ -\lambda_{2t}^*(S)\mathbf{u}_t(S) & \text{with probability} \quad 1 - q(t) \end{cases}$$

and $q(t) = \lambda_{2t}^*(S)/(\lambda_{1t}^*(S) + \lambda_{2t}^*(S))$.

The procedure ends at step $T(S)$, when $\boldsymbol{\pi}_{T(S)}(S)$ has only integer $(0-1)$ components.

B. An Unbiased Variance Estimator for V_{NC}

From the identity

$$\begin{aligned} \sum_{k,l \in S_r : k \neq l} \frac{\Delta_{kl,r}}{\alpha_{kl,r}\alpha_{kl|G}} a_{kl} &= \sum_{u_i \in S_I} \frac{1}{\pi_{Ii}} \sum_{k,l \in u_i : k \neq l} \frac{\Delta_{kl,r}}{\alpha_{kl,r}} a_{kl} 1(k, l \in G_r) \\ &+ \sum_{u_i, u_j \in S_I : u_i \neq u_j} \frac{1}{\pi_{Iij}} \sum_{k \in u_i} \sum_{l \in u_j} \frac{\Delta_{kl,r}}{\alpha_{kl,r}} a_{kl} 1(k, l \in G_r), \end{aligned} \tag{29}$$

and assuming all $\alpha_{kl,r}$ and $\alpha_{kl|G_r}$ to be strictly positive, we obtain

$$\begin{aligned} E\left(\sum_{k,l \in S_r : k \neq l} \frac{\Delta_{kl,r}}{\alpha_{kl,r}\alpha_{kl|G}} a_{kl} \middle| G_1, \ldots, G_R \right) &= \sum_{u_i \in U_I} \sum_{k,l \in u_i : k \neq l} \frac{\Delta_{kl,r}}{\alpha_{kl,r}} a_{kl} 1(k, l \in G_r) \\ + \sum_{u_i, u_j \in U_I : u_i \neq u_j} \sum_{k \in u_i} \sum_{l \in u_j} \frac{\Delta_{kl,r}}{\alpha_{kl,r}} a_{kl} 1(k, l \in G_r) &= \sum_{k,l \in U : k \neq l} \frac{\Delta_{kl,r}}{\alpha_{kl,r}} a_{kl} 1(k, l \in G_r), \end{aligned}$$

which leads to

$$\begin{aligned} E\left(\sum_{k,l \in S_r : k \neq l} \frac{\Delta_{kl,r}}{\alpha_{kl,r}\alpha_{kl|G}} a_{kl} \right) &= EE\left(\sum_{k,l \in S_r : k \neq l} \frac{\Delta_{kl,r}}{\alpha_{kl,r}\alpha_{kl|G}} a_{kl} \middle| G_1, \ldots, G_R \right) \\ &= E\left(\sum_{k,l \in U : k \neq l} \frac{\Delta_{kl,r}}{\alpha_{kl,r}} a_{kl} 1(k, l \in G_r) \right) = \sum_{k,l \in U : k \neq l} \Delta_{kl,r} a_{kl}. \end{aligned}$$

The result is thus obtained by plugging

$$a_{kl} = \left(\frac{y_k}{\alpha_{kr}} - \frac{y_l}{\alpha_{lr}} \right)^2$$

into (29).

7. References

Berlemont, B., Christine, M., and Faivre, S. (2009). The French New Master Sample 2009: Building Fresh Annual Sampling Frames for Household Surveys Based on the New Annual Census. New Techniques and Technologies for Statistics, Brussels.

Bourdalle, G., Christine, M., and Wilms, L. (2000). Échantillons maître et emploi. Série INSEE Méthodes, Paris, France, 21, 139–173. [In French]

Breidt, F.J. and Chauvet, G. (2011). Improved Variance Estimation for Balanced Samples Drawn via the Cube Method. Journal of Statistical Planning and Inference, 141, 479–487.

Chauvet, G., Bonnéry, D., and Deville, J.-C. (2011). Optimal Inclusion Probabilities for Balanced Sampling. Journal of Statistical Planning and Inference, 141, 984–994.

Chauvet, G. and Tillé, Y. (2005). Fast SAS Macros for Balancing Samples: User's Guide. Technical Report, University of Neuchâtel.

Christine, M. and Faivre, S. (2009). The French New Master Sample 2009: Building Fresh Annual Sampling Frames for Household Surveys Based on the New Annual Census. Presentation at the New Techniques and Technologies for Statistics (NTTS), Brussels, Belgium.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Deville, J.-C. and Tillé, Y. (2004). Efficient Balanced Sampling: The Cube Method. Biometrika, 91, 893–912.

Deville, J.-C. and Tillé, Y. (2005). Variance Approximation under Balanced Sampling. Journal of Statistical Planning and Inference, 128, 569–591.

Fattorini, L. (2006). Applying the Horvitz-Thompson Criterion in Complex Designs: A Computer Intensive Perspective for Estimating Inclusion Probabilities. Biometrika, 93, 269–278.

Fuller, W.A. (2009). Some Design Properties of a Rejective Sampling Procedure. Biometrika, 96, 933–944.

Godinot, A. (2005). Pour comprendre le Recensement de la population. Séries Insee Méthodes, Hors Série. [In French]

R Development Core Team (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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

Thompson, M. and Wu, C. (2008). Simulation-based Randomized Systematic pps Sampling under Substitution of Units. Survey Methodology, 34, 3–11.

Tillé, Y. (2006). Sampling Algorithms. Springer Series in Statistics. New York: Springer.

Tillé, Y. (2010). Balanced Sampling by Means of the Cube Method. Presentation to the International Statistical Seminar of EUSTAT, Bilbao, Basque Country.

Tillé, Y. and Favre, A.-C. (2004). Coordination, Combination and Extension of Balanced Samples. Biometrika, 91, 913–927.

Tillé, Y. and Favre, A.-C. (2005). Optimal Allocation in Balanced Sampling. Statistics and Probability Letters, 74, 31–37.

Tillé, Y. and Matei, A. (2008). Sampling: Survey Sampling. R package version 2.0.

Yates, F. and Grundy, P. (1953). Selection without Replacement from Within Strata with Probability Proportional to Size. Journal of the Royal Statistical Society, Series B, 15, 253–261.

Received November 2009
Revised August 2011