INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2010-02

Optimal Inclusion Probabilities for Balanced Sampling

G. CHAUVET[1]
D. BONNÉRY[2]
J-C. DEVILLE[3]

Les documents de travail ne reflètent pas la position de l'INSEE et n'engagent que leurs auteurs.

Working papers do not reflect the position of INSEE but only the views of the authors.

[1] Ecole Nationale de la Statistique et de l'Analyse de l'Information (ENSAI) CREST, Campus de Ker Lann, 35170 Bruz, France. Email : chauvet@ensai.fr

[2] ENSAI-CREST, Campus de Ker Lann, 35170 Bruz, France.

[3] ENSAI- CREST, Campus de Ker Lann, 35170 Bruz, France.

Optimal inclusion probabilities for balanced sampling

July 26, 2010

Guillaume Chauvet
Ecole Nationale de la Statistique et de l'Analyse de l'Information
Laboratoire de Statistique d'Enquête

Daniel Bonnéry
Jean-Claude Deville
Laboratoire de Statistique d'Enquête

ABSTRACT

The use of balanced sampling at the design stage of a survey requires the knowledge of auxiliary information for any unit in the population. The samples selected are such that the Horvitz-Thompson estimators of the auxiliary variables match the known totals of those variables, resulting in a variance reduction. In this paper, a method for computing optimal inclusion probabilities for balanced sampling on given auxiliary variables is studied. We show that the method formerly suggested by Tillé and Favre (2005) enables the computation of inclusion probabilities that lead to a decrease in variance. Since the target optimal inclusion probabilities usually depend on the variable of interest, we propose to use estimates instead (e.g., arising from a previous wave of the survey). A limited simulation study suggests that our method performs well as compared to that suggested by Tillé and Favre (2005).

Keywords : Balanced sampling ; Calibration ; Cube method ; Fixed-point algorithm ; Variance approximation.

1 Introduction

A sampling design is said to be balanced if it leads to the selection of samples such that the Horvitz-Thompson estimators of the totals for auxiliary variables exactly match the known population totals. Many partial solutions were proposed for balanced sampling, before Deville and Tillé (2004) introduced the *cube method*. This sampling algorithm enables the selection of balanced samples with any number of balancing variables, and any prescribed set of inclusion probabilities. The algorithm has been programmed into a SAS macro (Chauvet and Tillé, 2006, 2007) and is also available in the R Sampling Package prepared by Matei and Tillé (2006).

Balanced sampling designs do not substitute for other classical and efficient sampling techniques, such as unequal probability sampling for selecting primary sampling units (PSUs) in household surveys, or stratification in business surveys. They may be thought of as a way to refine these techniques and obtain a variance reduction, if auxiliary information is available at the design stage. For example, the Cube method was used for the selection of the PSUs in the 1999 French Master Sample (Bourdalle et al., 2000) with balanced sampling on variables such as taxable net income and age groups. In this paper, we propose to compute optimal inclusion probabilities for balanced sampling designs by means of a fixed-point algorithm, previously suggested by Tillé and Favre (2005). Under some conditions on the set of balancing variables, we show that the resulting inclusion probabilities always lead to a reduction in variance of the Horvitz-Thompson estimator. Whereas several iterations of the fixed-point algorithm are usually needed to get the target inclusion probabilities, we note that the set of inclusion probabilities obtained after one iteration is close to the final one. Consequently, considering only one iteration appears as a good trade-off between accuracy and simplicity. A disadvantage of the studied method is that some knowledge on the variable of interest is required, since quantities depending on the variable of interest are needed for the fixed-point algorithm. If these quantities are unknown at the design stage, we propose to use estimates arising from another survey instead. Our simulation results suggest that the proposed method performs well, as compared to the approximation originally proposed by Tillé and Favre (2005).

The paper is organized as follows. The notation is defined in Section 2. The algorithm for computing optimal inclusion probabilities is described in Section 3, and its properties are discussed. A limited simulation study is proposed in Section 4. Our main conclusions are drawn in Section 5.

2 Notation and Balanced Sampling

Let U denote a finite labeled population of size N. Let S denote a random sample selected in U by means of a sampling design $p(\cdot)$. Let π_k denote the inclusion probability of unit k, that is, the probability for unit k to be included in the sample S. Let π_{kl} denote the probability for distinct units k and l to be jointly in the sample. We note $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k, \ldots, \pi_N)'$ for the vector of inclusion probabilities. We assume that $\sum_{k \in U} \pi_k = n$, where n denotes the given average sample size.

Let y denote some variable of interest. In this paper, we are interested in estimating the population total $t_y = \sum_{k \in U} y_k$. The Horvitz-Thompson (HT) estimator is given by

$$\hat{t}_{y\pi} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k = \sum_{k \in U} d_k y_k I_k,$$

where $I_k = 1$ if unit k has been selected in the sample and 0 otherwise, and $d_k = 1/\pi_k$ denotes the design weight. This is a design-unbiased estimator for the total t_y. We look for a vector $\boldsymbol{\pi}$ of inclusion probabilities that minimizes, in some sense, the variance of the HT estimator. This variance is given by the so-called Horvitz-Thompson (1952) formula :

$$V(\hat{t}_{y\pi}) = \sum_{k,l \in U} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} (\pi_{kl} - \pi_k \pi_l). \tag{1}$$

We assume that a vector $\mathbf{x}_k = (x_{1k}, \cdots, x_{qk})'$ of q auxiliary variables is known at the design stage for each unit k in the population. The variables $\mathbf{x}_k$ are assumed without loss of generality to be linearly independent. The sampling design may be improved by means of the cube method (Deville and Tillé, 2004) which enables the selection of balanced samples. The sampling design $p(\cdot)$ is said to be balanced on variables $\mathbf{x}$ if the equations

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

hold exactly, where $t_{\mathbf{x}}$ gives the (vector) population total of variables $\mathbf{x}_k$. That is, the HT-estimators exactly match the known population totals. Consequently, the variance of the HT-estimator is zero for the *balancing variables.* We assume that the variables $\mathbf{x}$ are fixed prior to computing inclusion probabilities. That is, we assume that the balancing variables do not depend on the $\boldsymbol{\pi}$-vector of inclusion probabilities. Note that our set-up does not *a priori* cover the situation when a variable

$$x_{ak} = \pi_k 1(k \in U_1) \tag{3}$$

belongs to the balancing variables, where U_1 is a subset of units in U, and $1(k \in U_1)$ equals 1 if k belongs to the domain U_1, and 0 otherwise. Balancing on variable x_{ak} is equivalent to impose a condition of fixed size in domain U_1. For example, the condition of global fixed sample size is met if $x_{ak} = \pi_k$, that is, if $U_1 = U$. It is shown in Section 3.2 that this assumption may be partially relaxed.

As an exact balanced sample may usually not be found, the cube method enables the selection of approximately balanced samples. The algorithm may be split into two phases, called the *flight phase* and the *landing phase*. At each step of the flight phase, one unit is either selected in the sample or definitely rejected. The result of the flight phase is given by a vector $\boldsymbol{\pi}^* = (\pi_1^*, \cdots, \pi_k^*, \cdots, \pi_N^*)'$, where π_k^* equals 1 if unit k has been selected in the sample, 0 if unit k has been rejected from the sample, and lies between 0 and 1 otherwise. At the end of the flight phase, the balancing equations are exactly respected. That is,

$$\sum_{k \in U} \frac{\mathbf{x}_k}{\pi_k} \pi_k^* = \sum_{k \in U} \mathbf{x}_k. \tag{4}$$

In the case where some units are neither selected nor rejected after the flight phase, the landing phase consists in defining a sampling design among the remaining units, so that the inclusion probabilities are exactly respected and the variance due to the landing phase is minimized. Let $I = (I_1, \cdots, I_k, \cdots, I_N)'$ be the vector that gives the result of the landing phase. Then

$$E_p(I) = \boldsymbol{\pi} \tag{5}$$

and

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} \simeq \sum_{k \in U} \mathbf{x}_k, \tag{6}$$

where $E_p(\cdot)$ denotes the expectation with respect to the sampling design. Equation (5) states that the inclusion probabilities are exactly respected. Equation (6) implies that the sample is only approximately balanced, as the HT-estimator $\hat{t}_{\mathbf{x}\pi}$ usually does not exactly match the real total $t_{\mathbf{x}}$. If the sample is not exactly balanced, the sampling weights may be adjusted by means of *calibration techniques* (Deville and Särndal, 1992). The resulting calibration estimator of t_y is given by

$$\hat{t}_{yw} = \sum_{k \in S} d_k F(\lambda' \mathbf{x}_k) y_k, \tag{7}$$

where $F(\cdot)$ denotes the calibration function and λ is a q-vector of Lagrange multipliers. A special case of (7) is obtained under the linear function $F(u) = 1 + u$ which leads to the generalized regression estimator

$$\hat{t}_{y,greg} = \sum_{k \in S} w_k y_k, \tag{8}$$

where $w_k = d_k \left[1 + \left(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi} \right)' \hat{T}^{-1} \mathbf{x}_k \right]$ denotes the calibrated weight, with $\hat{T} = \sum\limits_{k \in S} d_k \mathbf{x}_k \mathbf{x}_k'$. Deville and Tillé (2004, section 8) give a short comparison of balanced sampling and calibration. They advocate for their joint use, since balanced sampling enables a reduction in the variability of the final weights, while calibration enables to match the known totals exactly.

A variance approximation is also provided by Deville and Tillé (2005). They suppose that the sampling design is exactly balanced, and performed with maximum entropy among sampling designs balanced on the same balancing variables, with the same inclusion probabilities. Then, under an additional assumption of asymptotic normality of the multivariate HT-estimator under Poisson sampling, they derive the following variance approximation :

$$V_{app}(\hat{t}_{y\pi}) = \frac{N}{N-q}\sum_{k\in U} b(\pi_k)\left(y_k - y_k^*(\pi)\right)^2, \tag{9}$$

where q denotes the number of balancing variables, $b(\pi_k) = 1/\pi_k - 1$ and $y_k^*(\boldsymbol{\pi}) = \mathbf{x}'_k\ \beta(\boldsymbol{\pi})$ is a weighted prediction of y_k obtained with the balancing variables, with

$$\beta(\pi) = \left(\sum_{l\in U} b(\pi_l)\mathbf{x}_l\mathbf{x}'{}_l\right)^{-1} \sum_{l\in U} b(\pi_l)\mathbf{x}_l y_l$$

Other slightly different approximations are proposed in Deville and Tillé (2005), but their simulation results suggest that approximation (9) performs well among variance approximations that may be computed in the case of any set of inclusion probabilities.

3 Optimal Allocation for Balanced Sampling

In many cases, inclusion probabilities are fixed and chosen to be proportional to an auxiliary variable known for any unit in the population at the design stage. Unequal probability sampling is then an efficient sampling design for estimating the total t_y if the variable of interest y is approximately proportional to this auxiliary variable. However, if some information on variable y is known at the design stage, it may be of interest to look for inclusion probabilities that minimize, at least approximately, the variance of the HT-estimator $\hat{t}_{y\pi}$. In what follows, section 3.1 mainly consists in a recall of Tillé and Favre (2005), apart from equation (12) which was only stated by these authors, and for which we give an explicit proof.

3.1 Optimal allocation for an approximation of the variance

An optimal vector $\boldsymbol{\pi}$ of inclusion probabilities should minimize the variance given in formula (1), under the constraints that

$$0 \leq \pi_k \leq 1 \text{ for any unit } k \in U \tag{10}$$

and

$$\sum_{k \in U} \pi_k = n. \tag{11}$$

Unfortunately, the variance formula (1) depends on second-order inclusion probabilities, and the link between the first and the second-order inclusion probabilities is intricate in case of balanced sampling, even in particular cases ; see Chen et al. (1994); Deville (2000); Matei and Tillé (2005) for the special case of balanced sampling on the sample size with maximum entropy, also denominated in the literature as rejective sampling (Hájek, 1964).

Following Tillé and Favre (2005), we thus propose to minimize the variance approximation (9) instead. This leads to the *Approximated Optimization Problem* (AOP) : seek for inclusion probabilities that minimize (9), under the constraints (10) and (11). The solution to this problem satisfies the system of equations

$$\pi_l = n \frac{|y_l - y_l^*(\pi)|}{\sum_{m \in U} |y_m - y_m^*(\pi)|} \text{ for any } l \in U, \tag{12}$$

where $y_l^*(\pi) = \mathbf{x}_l' \Bigg(\sum_{m \in U} b_m \mathbf{x}_m \mathbf{x}'_m \Bigg)^{-1} \sum_{m \in U} b_m \mathbf{x}_m y_m$ and $b_l = 1/\pi_l - 1$. The proof is given in Appendix A. This system of equations may not be used to compute directly the optimal inclusion probabilities, since both parts of each equation depend on $\boldsymbol{\pi}$. Intuitively, this formula states that if the absolute value of the residual $|e_k| = |y_k - y_k^*(\boldsymbol{\pi})|$ is large, the inclusion probability of unit k should be large, too. Conversely, a small inclusion probability should be associated with a small residual. In other words, there is no need to give large inclusion probabilities for units k such that y_k may be well predicted by the balancing variables, and attention should be paid to the remaining units instead.

A fixed-point algorithm may be used to compute the inclusion probabilities associated with formula (12), but the value of the variable of interest y is required for any unit in the population, and such detailed information is unknown at the design stage. This first set of inclusion probabilities is thus difficult to compute in practice.

3.2 Generalization of the Approximated Optimization Problem

To overcome this difficulty, we propose a generalization of this optimization problem. Assume that a categorical variable z is known. This may be one of the balancing variables or any additional variable available at the design stage for any unit in the population. This variable defines a partition of the population into J non-overlapping subsets $U_1, \ldots, U_J$ of sizes $N_1, \ldots, N_J$, respectively, where J denotes the number of categories of the variable. Then we impose that the target inclusion probabilities satisfy the following system of equations :

$$\pi_k = \alpha_j \text{ for any unit } k \in U_j,\ j = 1, \ldots, J. \tag{13}$$

That is, inclusion probabilities are assumed to be equal inside each subset U_j. The variance approximation given in formula (9) may then be alternatively written as

$$V_{app}(\hat{t}_{y\pi}) \equiv V(\boldsymbol{\alpha}) = \frac{N}{N-q} \sum_{j=1}^{J} b(\alpha_j) \sum_{k \in U_j} (y_k - \tilde{y}_k(\boldsymbol{\alpha}))^2, \tag{14}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)'$, $b(\alpha_j) = 1/\alpha_j - 1$ and

$$\tilde{y}_k(\boldsymbol{\alpha}) = \mathbf{x}_k' (\sum_{j=1}^{J} b(\alpha_j) \mathbf{A}_j)^{-1} \sum_{j=1}^{J} b(\alpha_j) \mathbf{c}_{1j}(y)$$

with $\mathbf{A}_j = \sum_{k \in U_j} \mathbf{x}_k \mathbf{x}_k'$ and $\mathbf{c}_{1j}(y) = \sum_{k \in U_j} \mathbf{x}_k y_k$. The *General Approximated Optimization Problem* (GAOP) may then be described as follows : find the $J \times 1$ vector $\boldsymbol{\alpha}$ that minimizes (14) under the constraints (10), (11) and (13).

Such a vector satisfies the system of equations

$$\alpha_j = n \frac{\sigma_j(\boldsymbol{\alpha})}{\sum_{i=1}^J N_i \sigma_i(\boldsymbol{\alpha})}, \tag{15}$$

where

$$\sigma_j(\boldsymbol{\alpha}) = \frac{1}{N_j} \sum_{k \in U_j} (y_k - \tilde{y}_k(\boldsymbol{\alpha}))^2. \tag{16}$$

The proof is similar to that of (12), and is thus omitted. Note that the AOP is a special case of our setting, obtained when $J = N$. In practice, the domains associated to the variable z should be chosen so that the quantities needed for the computation of the inclusion probabilities may be either known or accurately estimated from an external source, see section 3.3.

Once again, we note that the formula (15) may not be directly used to compute optimal inclusion probabilities since both parts in (15) depend on the unknown $\boldsymbol{\alpha}$. The fixed-point Algorithm 1 may be used instead.

Algorithm 1 : Fixed-point algorithm to compute optimal inclusion probabilities

```
1. Initialize with any vector $\boldsymbol{\alpha}^0 = (\alpha_1^0, \ldots, \alpha_J^0)'$.
2. At step $t$, compute $\boldsymbol{\alpha}^t = (\alpha_1^t, \ldots, \alpha_J^t)'$ such that
       $\alpha_j^t = n \frac{\sigma_j(\boldsymbol{\alpha}^{t-1})}{\sum_{i=1}^J N_i \sigma_i(\boldsymbol{\alpha}^{t-1})}$ for any $j = 1, \ldots, J$.
3. The procedure ends at step $T$ when $Max_j \|\alpha_j^t - \alpha_j^{t-1}\|$ is lower than a
   pre-specified bound $\epsilon$.
```

The following result states that Algorithm 1 always lead to a reduction in variance, as compared to the variance associated with the original $\boldsymbol{\alpha}^0$-vector.

Theorem 1 *At any step t of the fixed point Algorithm 1, $V(\boldsymbol{\alpha}^t) \leq V(\boldsymbol{\alpha}^{t-1})$.*

The proof is given in Appendix B. As a consequence, the sequence $(\boldsymbol{\alpha}^t)_{t\in\mathbb{N}}$ tends to a local minimum, and the approximated variance is always improved if the inclusion probabilities are given by the fixed-point algorithm. With the simulations performed and a bound of $\epsilon = 10^{-6}$, very few iterations are required, so that $\boldsymbol{\alpha}^1$ provides a good approximation of the target vector of inclusion probabilities.

We now consider the problem of the choice of the categorical variable z whose categories are used to partition the population into domains with equal probabilities inside. Both the AOP and the GAOP should give comparable results if the absolute value of the residuals $|e_k| = |y_k - y_k^*(\boldsymbol{\pi})|$ are approximately equal inside domains $U_1, \ldots, U_J$. That is, the population U should be sorted according to the $|e_k|$ variable, and the domains separated by the fractiles of this variable. Since these residuals are practically unknown at the design-stage, an alternative consists in using the available auxiliary information. For example, qualitative variables used in the vector $\mathbf{x}_k$ of balancing variables could also be used to define the domains. Also, we previously assumed that the balancing variables did not depend on the inclusion probabilities, and in particular that no constraint on fixed size was involved in the balancing problem. This latter assumption may be relaxed if the domains inside which fixed sample size is required are used as the domains $U_1, \ldots, U_J$ in the GAOP. Let us suppose that the categorical variable z defining the domains belongs to the balancing variables. The corresponding balancing equations may be written as

$$\sum_{k \in S} \frac{1(k \in U_j)}{\pi_k} = \sum_{k \in U} 1(k \in U_j) \tag{17}$$

for any domain $U_j, j = 1, \ldots, J$, and the joint application of equations (13) and (17) leads to

$$n(S_j) = \alpha_j N_j, \tag{18}$$

where $n(S_j)$ denotes the size of the sub-sample $S_j = S \cap U_j$. The set of equations (18) impose that the sample size is fixed inside each domain U_j, but since $\alpha_j N_j$ may not be an integer the balancing constraints (18) will usually be respected to within about one unit. Note that the summation of equations (18) leads to

$$\begin{array}{rcccc} n(S) & = & \sum_{j=1}^{J} n(S_j) & = & \sum_{j=1}^{J} \alpha_j N_j \\ & = & \sum_{k \in U} \pi_k & = & n \end{array}$$

by application of equation (11), so that if z belongs to the balancing variables, the condition of global fixed sample size will be exactly respected.

3.3 Practical implementation of the Optimization Problem

Once again, we note that some knowledge about the variable of interest y is needed in the fixed-point algorithm. More specifically, the knowledge of $\mathbf{A}_j = \sum_{k \in U_j} \mathbf{x}_k \mathbf{x}_k'$, $\mathbf{c}_{1j}(y) = \sum_{k \in U_j} \mathbf{x}_k y_k$ and $\mathbf{c}_{2j}(y) = \sum_{k \in U_j} y_k^2$ is needed for any subset U_j. Though some of these quantities are usually unknown at the design stage, they may be replaced by estimated quantities. This is a common practice to take advantage of accurate estimated totals to improve the estimators arising from a survey, see Berger et al. (2009). For example, these estimated totals may be obtained from a previous wave or occasion of the survey, or from a larger survey ; household surveys conducted in France are usually calibrated on estimates arising from the Labour Force Survey. Let us suppose that another sample S^p has been selected in U with inclusion probabilities $\boldsymbol{\pi}^p = (\pi_1^p, \ldots, \pi_k^p, \ldots, \pi_N^p)'$. Let $\hat{\sigma}_j(\alpha)$ be obtained from (16) by replacing $\mathbf{A}_j$, $\mathbf{c}_{1j}(y)$ and $\mathbf{c}_{2j}(y)$ with $\hat{\mathbf{A}}_j^p = \sum_{k \in S_j^p} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k^p}$, $\hat{\mathbf{c}}_{1j}^p(y) = \sum_{k \in S_j^p} \frac{\mathbf{x}_k y_k}{\pi_k^p}$ and $\hat{\mathbf{c}}_{2j}^p(y) = \sum_{k \in S_j^p} \frac{y_k^2}{\pi_k^p}$ respectively, where $S_j^p = S^p \cap U_j$. Algorithm 2 may then be used to compute approximately optimal inclusion probabilities, that we denote

$$\hat{\boldsymbol{\pi}}^T = (\hat{\pi}_1^T, \ldots, \hat{\pi}_k^T, \ldots, \hat{\pi}_N^T)' \tag{19}$$

where $\hat{\pi}_k^T = \hat{\alpha}_j^T$ for any unit $k \in U_j,\ j = 1, \ldots, J$. Since the exact quantities $\mathbf{A}_j$, $\mathbf{c}_{1j}(y)$ and $\mathbf{c}_{2j}(y)$ are not used in Algorithm 2, the computed inclusion probabilities do not necessarily lead to an optimal solution. However, the use of unbiased estimators $\hat{\mathbf{A}}_j$, $\hat{\mathbf{c}}_{1j}(y)$ and $\hat{\mathbf{c}}_{2j}(y)$ should lead to a strong reduction of the variance of the Horvitz-Thompson estimator, even if this variance is not minimized (see section 4.2).

We now briefly discuss the alternative solution proposed by Tillé and Favre (2005). For simplicity, we assume that the same variable of interest y and auxiliary variables $\mathbf{x}$ are collected in both the samples S^p and S. First, estimated residuals

$$\hat{e}_{k1} = y_k - \mathbf{x}_k' \hat{B}^p \tag{20}$$

Algorithm 2 : Fixed-point algorithm to compute approximately optimal inclusion probabilities

```
1. Initialize with any vector $\hat{\boldsymbol{\alpha}}^0 = (\hat{\alpha}_1^0, \ldots, \hat{\alpha}_J^0)'$.
2. At step $t$, compute $\hat{\boldsymbol{\alpha}}^t = (\hat{\alpha}_1^t, \ldots, \hat{\alpha}_J^t)'$ such that
       $\hat{\alpha}_j^t = n \frac{\hat{\sigma}_j(\hat{\boldsymbol{\alpha}}^{t-1})}{\sum_{i=1}^J N_i \hat{\sigma}_i(\hat{\boldsymbol{\alpha}}^{t-1})}$ for any $j = 1, \ldots, J.$
3. The procedure ends at step $T$ when $Max_j \|\hat{\alpha}_j^t - \hat{\alpha}_j^{t-1}\|$ is lower than a
   pre-specified bound $\epsilon$.
```

are computed for units $k \in S^p$, where

$$\hat{B}^p = \left(\sum_{k \in S^p} \frac{1-\pi_k^p}{(\pi_k^p)^2} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in S^p} \frac{1-\pi_k^p}{(\pi_k^p)^2} \mathbf{x}_k y_k.$$

Then, a linear model

$$|\hat{e}_{k1}|^2 = \mathbf{x}_k' \boldsymbol{\psi} + \epsilon_k \tag{21}$$

is postulated to predict the link between the square residuals and the auxiliary variables, where $\boldsymbol{\psi}$ is a q-vector of unknown parameters and the ϵ_k's are residuals. An estimator $\hat{\boldsymbol{\psi}}^p$ of the vector $\boldsymbol{\psi}$ is obtained from sample S^p, to get estimated square residuals

$$|\hat{e}_{k2}|^2 = \mathbf{x}_k' \hat{\boldsymbol{\psi}}^p \tag{22}$$

for any unit $k \in U$. Finally, the optimal inclusion probabilities are estimated by

$$\hat{\pi}_k^{TF} = n \frac{|\hat{e}_{k2}|}{\sum_{l \in U} |\hat{e}_{l2}|}. \tag{23}$$

If the quantities computed in (23) are larger than 1, Tillé and Favre (2005) propose to set the concerned inclusion probabilities to 1, while the other inclusion probabilities are recalculated. The method proposed by Tillé and Favre is less computer-intensive than the method that we propose, since no fixed-point algorithm is required for the computation of the inclusion probabilities. However, formula (22) may lead to negative estimated square residuals for some units in U. In that case, the associated inclusion probabilities may be set to 0, which results in biased HT-estimators. Moreover, the quality of the prediction given in (22) highly depends on the predictive power of the auxiliary variables $\mathbf{x}_k$ for the residuals. If this predictive power is poor, the estimated inclusion probabilities given in (23) may fall far from the optimal probabilities, resulting in a possible loss of efficiency. The method proposed by Tillé and Favre (2005) as well as the proposed method are compared in section 4 into a small simulation study.

4 A simulation study

We conducted a limited simulation study to test the performance of the procedures described in section 3. We first generated a finite population of size $N = 1\ 000$ containing 6 variables : three variables of interest y_1, y_2 and y_3 and three auxiliary variables x_0, x_1 and x_2. First, the values of the variable x_0 were generated independently from a uniform distribution. The population U was divided into four groups $U_1, \ldots, U_4$ according to the quartiles of the x_0-values, and the population x_{1k} and x_{2k} were generated as

$$x_{1k} = \begin{cases} 1 & \text{if } k \in U_1 \cup U_2 \\ 2 & \text{otherwise} \end{cases}$$

and

$$x_{2k} = \begin{cases} 1 & \text{if } k \in U_1 \cup U_3 \\ 2 & \text{otherwise} \end{cases}$$

Given the values of these auxiliary variables, the $y_1,\ y_2$ and y_3-values were generated inside each group U_j according to the model

$$y_{hk} = \phi_{hj} + \eta_{jk}, h = 1, \ldots, 3. \tag{24}$$

The η_{jk}'s were generated according to a normal distribution with mean 0 and variance σ_j^2. The vector of model parameters $\phi_h = (\phi_{h1}, \phi_{h2}, \phi_{h3}, \phi_{h4})$ was set to $\phi_1 = (0.5, 0.5, 1.5, 1.5)$ for variable y_1, $\phi_2 = (0.5, 1.5, 0.5, 1.5)$ for variable y_2 and $\phi_3 = (0.2, 0.75, 1.25, 2.0)$ for variable y_3. That is, y_1 and y_2 are related to the auxiliary variables x_1 and x_2 respectively, whereas the variable y_3 is related to the interaction of variables x_1 and x_2. This last variable of interest is meant to evaluate (to some extent) the performance of the computed inclusion probabilities when the auxiliary information used is not fully adequate. We used two possible values for the vector $\sigma = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)$, namely $\sigma^{(1)} = (0.2, 0.3, 0.4, 0.5)$ and $\sigma^{(2)} = (0.4, 0.6, 0.8, 1.0)$.

4.1 Simulation 1 : optimal inclusion probabilities

We first assume that, for each variable of interest $y_h,\ h = 1, \ldots, 3$, the needed quantities $\mathbf{A}_j$, $\mathbf{c}_{1j}(y_h)$ and $\mathbf{c}_{2j}(y_h)$ are exactly known. These quantities are given in Table 1.

The inclusion probabilities are assumed to be equal inside each group U_j. For each variable of interest $y_h,\ h = 1, \ldots, 3$, we note α_{hj} for the common inclusion probability for units in U_j and $\boldsymbol{\alpha}_h = (\alpha_{h1}, \alpha_{h2}, \alpha_{h3}, \alpha_{h4})'$. Algorithm 1 is

Table 1 – Exact quantities needed for the computation of optimal inclusion probabilities with Algorithm 1, for the vectors $\sigma^{(1)}$ and $\sigma^{(2)}$

		U_1	U_2	U_3	U_4
$\mathbf{A}_j$		$\begin{pmatrix} 250 & 250 \\ 250 & 250 \end{pmatrix}$	$\begin{pmatrix} 250 & 500 \\ 500 & 1000 \end{pmatrix}$	$\begin{pmatrix} 1000 & 500 \\ 500 & 250 \end{pmatrix}$	$\begin{pmatrix} 1000 & 1000 \\ 1000 & 1000 \end{pmatrix}$
		$\sigma^{(1)}$			
$\mathbf{c}'_{1j}(\cdot)$	y_1	$(122.16, 122.16)$	$(124.53, 249.05)$	$(752.32, 376.16)$	$(754.93, 754.93)$
	y_2	$(129.39, 129.39)$	$(376.65, 753.31)$	$(236.70, 118.35)$	$(765.42, 765.42)$
	y_3	$(63.75, 63.75)$	$(188.15, 376.30)$	$(620.97, 310.49)$	$(1004.64, 1004.64)$
$\mathbf{c}_{2j}(\cdot)$	y_1	68.68	80.71	609.69	632.32
	y_2	76.41	587.50	99.26	649.67
	y_3	26.56	161.68	425.57	1059.42
		$\sigma^{(2)}$			
$\mathbf{c}'_{1j}(\cdot)$	y_1	$(119.31, 119.31)$	$(124.05, 248.11)$	$(754.65, 377.32)$	$(759.85, 759.85)$
	y_2	$(133.78, 133.78)$	$(378.31, 756.62)$	$(223.40, 111.70)$	$(780.83, 780.83)$
	y_3	$(65.00, 65.00)$	$(188.80, 377.60)$	$(616.94, 308.47)$	$(1009.29, 1009.29)$
$\mathbf{c}_{2j}(\cdot)$	y_1	92.92	136.29	744.27	827.01
	y_2	109.35	652.56	222.84	864.91
	y_3	58.11	222.91	540.47	1219.11

initialized with equal probabilities $\boldsymbol{\alpha}_h^0 = (0.1, 0.1, 0.1, 0.1)'$ (EQUAL). Also, two other sets of inclusion probabilities are computed : (i) probabilities $\boldsymbol{\alpha}_h^1$ obtained after the first step (FSTEP) of Algorithm 1 and (ii) probabilities $\boldsymbol{\alpha}_h^T$ obtained at the end (LSTEP) of Algorithm 1. The corresponding $\boldsymbol{\alpha}$ vectors are presented in Table 2. In line with formula (12), we note that the optimal inclusion probabilities lead to larger sample sizes in domains where the variable of interest is highly dispersed, or more precisely in domains where the balancing variables have a lower explanatory power. The values taken by the variance approximation in formula (9) for the three different sets of inclusion probabilities are presented in Table 3, as well as the totals of the variables of interest. As expected, the approximated variance obtained with the final inclusion probabilities is systematically lower than the approximated variance obtained with the initial equal inclusion probabilities. The FSTEP and LSTEP inclusion probabilities give almost identical approximated variance, since the two sets of inclusion probabilities are very close in any case considered in the simulation, see Table 2. Though the results obtained after the first step may depend on the initial $\boldsymbol{\alpha}_h^0$, the vector $\boldsymbol{\alpha}_h^1$ may be expected to give a good compromise between variance reduction and low algorithmic complexity.

Table 2 – Three sets of inclusion probabilities obtained with the fixed-point algorithm for three variables of interest, for the vectors $\sigma^{(1)}$ and $\sigma^{(2)}$

		$\sigma^{(1)}$	$\sigma^{(2)}$
y_1	EQUAL	$(0.1, 0.1, 0.1, 0.1)$	$(0.1, 0.1, 0.1, 0.1)$
	FSTEP	$(0.055, 0.079, 0.121, 0.145)$	$(0.055, 0.079, 0.121, 0.145)$
	LSTEP	$(0.055, 0.079, 0.121, 0.145)$	$(0.055, 0.079, 0.121, 0.145)$
y_2	EQUAL	$(0.1, 0.1, 0.1, 0.1)$	$(0.1, 0.1, 0.1, 0.1)$
	FSTEP	$(0.056, 0.081, 0.119, 0.144)$	$(0.056, 0.081, 0.119, 0.144)$
	LSTEP	$(0.056, 0.081, 0.119, 0.144)$	$(0.056, 0.081, 0.119, 0.144)$
y_3	EQUAL	$(0.1, 0.1, 0.1, 0.1)$	$(0.1, 0.1, 0.1, 0.1)$
	FSTEP	$(0.063, 0.085, 0.119, 0.133)$	$(0.061, 0.085, 0.120, 0.134)$
	LSTEP	$(0.061, 0.085, 0.120, 0.134)$	$(0.061, 0.085, 0.120, 0.134)$

Table 3 – Total of the variables of interest and variance approximation for three sets of inclusion probabilities

		$\sigma^{(1)}$			$\sigma^{(2)}$		
		y_1	y_2	y_3	y_1	y_2	y_3
Total		1 000.31	1 007.10	1 064.71	1 000.62	1 014.21	1 066.92
Variance	EQUAL	1 269.97	1 347.95	1 277.54	5 079.87	5 391.80	5 004.06
Approximation	FSTEP	1 129.58	1 189.98	1 149.92	4 518.31	4 759.92	4 494.06
	LSTEP	1 129.58	1 189.98	1 149.69	4 518.31	4 759.92	4 493.99

The formula (9) which is minimized to compute optimal inclusion probabilities only gives an approximation for the true variance, under conditions that may fail in practice. For example, Deville and Tillé (2005) assume that the sampling design is exactly balanced, which is often unlikely to occur. Thus, it seems of interest to compare the performances of the different sets of inclusion probabilities with respect to the exact variance. We selected $B = 10\ 000$ samples of size $n = 100$, by balanced sampling on variables x_1 and x_2 by means of the Cube method, with the procedures EQUAL and LSTEP. Under each procedure, we computed the calibrated after balancing estimator, given by (7). As a measure of variability of an estimator , we used the Monte Carlo Mean Square Error (MSE) given by

$$MSE_{MC}(\hat{t}_{yw}) = \frac{1}{10\ 000} \sum_{b=1}^{B} \left(\hat{t}_{yw}(S_b) - t_y\right)^2, \tag{25}$$

where $\hat{t}_{yw}(S_b)$ denotes the estimator $\hat{t}_{yw}$ in the b-th sample S_b, $b = 1, \ldots, 10\ 000$. Let $\hat{t}_{yw}^{(EQUAL)}$ and $\hat{t}_{yw}^{(LSTEP)}$ denote the estimator $\hat{t}_{yw}$ under EQUAL and LSTEP, respectively. In order to compare the relative variability of the estimators, using $\hat{t}_{yw}^{(EQUAL)}$ as the reference, we used the following measure :

$$RE = \frac{MSE_{MC}(\hat{t}_{yw}^{(LSTEP)})}{MSE_{MC}(\hat{t}_{yw}^{(EQUAL)})}. \tag{26}$$

Table 4 shows the RE for the three variables. It is clear that the computed inclusion probabilities lead to a more efficient estimator in all the scenarios with a value of RE varying from 0.89 to 0.92.

Table 4 – Relative Efficiency of the optimal vector of inclusion probabilities

	y_1	y_2	y_3
$\sigma^{(1)}$	0.91	0.92	0.89
$\sigma^{(2)}$	0.89	0.90	0.92

4.2 Simulation 2 : approximately optimal inclusion probabilities

We conducted another simulation study to take into account the practical situation when the needed quantities $\mathbf{A}_j$, $\mathbf{c}_{1j}(y_h)$ and $\mathbf{c}_{2j}(y_h)$ are unknown. That is, the computation of optimal inclusion probabilities by means of Algorithm 1 requires some knowledge on the variable of interest y, that may not be available at the design stage. In that case, we assume that some information has been collected on a sample S^p, prior to the selection of the sample S. That is, a sample S^p is first selected in U, and the values of the variables of interest y_{hk} and of the auxiliary variables $\mathbf{x}_k$ are measured for any unit $k \in S^p$. The needed quantities are then replaced by unbiased estimates $\hat{\mathbf{A}}_j^p$, $\hat{\mathbf{c}}_{1j}^p(y_h)$ and $\hat{\mathbf{c}}_{2j}^p(y_h)$ (see section 3.3), and approximately optimal inclusion probabilities $\hat{\pi}_k^T$ given in (19) are obtained by means of Algorithm 2. The sample S is then selected by means of balanced sampling with inclusion probabilities $\hat{\pi}_k^T$. Alternatively, the method proposed by Tillé and Favre (2005) may be used instead of Algorithm 2 to obtain inclusion probabilities $\hat{\pi}_k^{TF}$ given by (23), and then to select the sample S.

We selected $B = 10,000$ samples S_b^p, $b = 1, \ldots, 10\ 000$ by simple random sampling of size $n_0 = 50$ (respectively, $n_0 = 100$). Then, several sets of inclusion probabilities are computed for any unit $k \in U$. The inclusion probabilities are equal inside each group U_j. For each sample S_b^p, Algorithm 2 is initialized with equal probabilities $\hat{\boldsymbol{\alpha}}_h^0 = (0.1, 0.1, 0.1, 0.1)'$ (EQUAL). Two other sets of inclusion probabilities are computed : (i) probabilities $\hat{\boldsymbol{\alpha}}_{bh}^T$ obtained at the end (APPROX) of Algorithm 2, and (ii) probabilities $\hat{\boldsymbol{\alpha}}_{bh}^{TF}$ associated to the method of Tillé and Favre (MODEL). Then, a sample $S_b,\ b = 1, \ldots, 10\ 000$ of size $n = 100$ is selected by balanced sampling on variables x_1 and x_2 by means of the Cube method, with the procedures EQUAL, APPROX and MODEL.

To compare the approximately optimal inclusion probabilities associated to the procedures APPROX and MODEL with the true, optimal inclusion probabilities associated to the LSTEP procedure (see section 4.1), we used the Monte Carlo Mean (MEAN), given by

$$MEAN_{MC}\left(\hat{\boldsymbol{\alpha}}_h^{(.)}\right) = \frac{1}{10\ 000} \sum_{b=1}^{B} \hat{\boldsymbol{\alpha}}_{bh}^{(.)}. \tag{27}$$

We present in Table 5 the Monte Carlo Mean obtained with APPROX and MODEL and a size of $n_0 = 50$ for the prior sample. The results obtained with $n_0 = 100$ were almost identical, and are thus omitted. Clearly, the Monte Carlo Bias associated to the proposed method is negligible so that APPROX may be expected to give results close to that of LSTEP. On the other hand, the Monte Carlo Bias associated to MODEL is non-negligible, except for the variable y_3, which may result in a loss of efficiency. To evaluate the performances of each procedure, we computed for each of them the calibrated after balancing estimator, given by (7). As a measure of variability of an estimator , we used the Monte Carlo Mean Square Error (MSE) given by equation (25) where $\hat{t}_{yw}(S_b)$ denotes the estimator $\hat{t}_{yw}$ in the b-th sample S_b, $b = 1, \ldots, 10\,000$. Let $\hat{t}_{yw}^{(EQUAL)}$, $\hat{t}_{yw}^{(APPROX)}$ and $\hat{t}_{yw}^{(MODEL)}$ denote the estimator $\hat{t}_{yw}$ under EQUAL, APPROX and MODEL, respectively. In order to compare the relative variability of the estimators, using $\hat{t}_{yw}^{(EQUAL)}$ as the reference, we used the following measure :

$$RE = \frac{MSE_{MC}(\hat{t}_{yw}^{(.)})}{MSE_{MC}(\hat{t}_{yw}^{(EQUAL)})}. \tag{28}$$

The results are presented in Table 6. Once again, we note that the inclusion probabilities computed with APPROX lead to a more efficient estimator than EQUAL, with values of RE ranging from 0.88 to 0.95. We note that the RE is closer to one when the sample size decreases. That is, a smaller size of the external survey used to estimate the needed quantities results in a loss of accuracy of the computed inclusion probabilities, as could be expected. Therefore, we advocate for the use of domains in which these needed quantities may be precisely estimated. Also, we note that MODEL gives quite poor results since it is outperformed by APPROX in all cases, and by EQUAL in 10 out of 12 cases.

5 Concluding remarks

In this paper, we studied the problem of computation of inclusion probabilities in the context of balanced sampling. We showed that, under some conditions on the vector of balancing variables, the fixed-point algorithm earlier suggested by Tillé and Favre (2005) to compute inclusion probabilities systematically leads to a decrease of the variance of the Horvitz-Thompson estimator. This algorithm requires that some quantities may be known from

Table 5 – Optimal inclusion probabilities given by Algorithm 1 and Monte Carlo Mean of the approximately optimal inclusion probabilities given by Algorithm 2 or by the method of Tillé and Favre for three variables of interest, obtained with $n_0 = 50$ for the vectors $\sigma^{(1)}$ and $\sigma^{(2)}$

		$\sigma^{(1)}$	$\sigma^{(2)}$
y_1	LSTEP	$(0.055, 0.079, 0.121, 0.145)$	$(0.055, 0.079, 0.121, 0.145)$
	APPROX	$(0.055, 0.079, 0.121, 0.144)$	$(0.055, 0.079, 0.121, 0.144)$
	MODEL	$(0.047, 0.084, 0.126, 0.143)$	$(0.047, 0.084, 0.126, 0.143)$
y_2	LSTEP	$(0.056, 0.081, 0.119, 0.144)$	$(0.056, 0.081, 0.119, 0.144)$
	APPROX	$(0.056, 0.081, 0.119, 0.144)$	$(0.056, 0.081, 0.119, 0.144)$
	MODEL	$(0.047, 0.086, 0.124, 0.143)$	$(0.047, 0.086, 0.124, 0.143)$
y_3	LSTEP	$(0.061, 0.085, 0.120, 0.134)$	$(0.061, 0.085, 0.120, 0.134)$
	APPROX	$(0.061, 0.085, 0.121, 0.134)$	$(0.061, 0.085, 0.120, 0.134)$
	MODEL	$(0.061, 0.086, 0.120, 0.133)$	$(0.060, 0.085, 0.121, 0.134)$

Table 6 – Relative Efficiency for two vectors of inclusion probabilities computed with respect to prior information known from a past survey

		$n_0 = 50$			$n_0 = 100$		
		y_1	y_2	y_3	y_1	y_2	y_3
$\sigma^{(1)}$	APPROX	0.93	0.95	0.95	0.88	0.91	0.89
	MODEL	1.13	1.20	1.31	1.00	1.04	0.98
$\sigma^{(2)}$	APPROX	0.93	0.89	0.94	0.92	0.88	0.90
	MODEL	1.13	1.22	1.17	1.03	1.01	0.96

an external source. If not, we proposed an alternative algorithm where the needed quantities are estimated. This situation is not uncommon in practice ; since most surveys are periodic, it may be of interest to take advantage of the previous waves of a survey. Results from a limited simulation study have shown that, even in the latter case, a significant decrease of the variance may be expected.

Further investigation on the matter is needed. First, the case where the balancing variables include some fixed-size constraints on domains is not covered by our set-up, if these domains do not coincide with those used in the GAOP. Such constraints are frequently needed, for example if a given level of precision is required for certain subdivisions of the population. Secondly, the approximation of variance of Deville and Tillé (2005) used to compute the inclusion probabilities is unlikely to hold if the assumption of maximum entropy is not satisfied. A practical way to increase the entropy of a sampling design is to sort the population randomly before the sampling. However, this preliminary randomization step is not systematically included in the sampling process. The case where the population is sorted with respect to some known auxiliary variable before balanced sampling is a matter for further research.

A Proof of equation (12)

To simplify the notation, we note $b(\pi_k) \equiv b_k$. First note that the optimization problem is equivalent to find the vector $\mathbf{b} = (b_1, \ldots, b_N)'$ that minimizes

$$W_0(\mathbf{b}) = \sum_{k \in U} b_k \left(y_k - y_k^0(\mathbf{b}) \right)^2$$

where $y_k^0(\mathbf{b}) = \mathbf{x}_k' \left(\sum_{l \in U} b_l \mathbf{x}_l \mathbf{x}_l' \right)^{-1} \sum_{l \in U} b_l \mathbf{x}_l y_l$, under the constraints :

$$b_k \geq 0 \text{ for any unit } k \in U \tag{29}$$

and

$$\sum_{k \in U} \frac{1}{b_k + 1} = n. \tag{30}$$

The partial derivative of $W_0(\mathbf{b})$, with respect to b_l, is equal to

$$\frac{\partial W_0(\mathbf{b})}{\partial b_l} = \left(y_l - y_l^0(\mathbf{b})\right)^2 + 2\sum_{k\in U} b_k \left(y_k - y_k^0(\mathbf{b})\right) \frac{\partial \left(y_k - y_k^0(\mathbf{b})\right)}{\partial b_l}. \qquad (31)$$

Since $y_k^0(\mathbf{b})$ may alternatively be written as $y_k^0(\mathbf{b}) = \mathbf{x}_k'\mathbf{A}(\mathbf{b})^{-1}\mathbf{c}(\mathbf{b})$, with $\mathbf{A}(\mathbf{b}) = \sum_{l\in U} b_l\mathbf{x}_l\mathbf{x}_l'$ and $\mathbf{c}(\mathbf{b}) = \sum_{l\in U} b_l\mathbf{x}_l y_l$, and since $\mathbf{x}_k$ does not depend on b_l, we have

$$\begin{array}{rcl}
\frac{\partial y_k^0(\mathbf{b})}{\partial b_l} & = & \mathbf{x}_k'\left(\frac{\partial(\mathbf{A}(\mathbf{b})^{-1})}{\partial b_l}\mathbf{c}(\mathbf{b}) + \mathbf{A}(\mathbf{b})^{-1}\frac{\partial \mathbf{c}(\mathbf{b})}{\partial b_l}\right)\\
 & = & \mathbf{x}_k'\left(-\mathbf{A}(\mathbf{b})^{-1}\frac{\partial \mathbf{A}(\mathbf{b})}{\partial b_l}\mathbf{A}(\mathbf{b})^{-1}\mathbf{c}(\mathbf{b}) + \mathbf{A}(\mathbf{b})^{-1}\frac{\partial \mathbf{c}(\mathbf{b})}{\partial b_l}\right)\\
 & = & \mathbf{x}_k'\mathbf{A}(\mathbf{b})^{-1}\mathbf{x}_l\left(y_l - \mathbf{x}_l'\mathbf{A}(\mathbf{b})^{-1}\mathbf{c}(\mathbf{b})\right).
\end{array}$$

By inserting this last expression into (31), we obtain

$$\begin{array}{rcl}
\frac{\partial W_0(\mathbf{b})}{\partial b_l} & = & \left(y_l - y_l^0(\mathbf{b})\right)^2\\
 & & -2\left[\sum_{k\in U} b_k(y_k - y_k^0(\mathbf{b}))\mathbf{x}_k'\right]\mathbf{A}(\mathbf{b})^{-1}\mathbf{x}_l\left(y_l - \mathbf{x}_l'\mathbf{A}(\mathbf{b})^{-1}\mathbf{c}(\mathbf{b})\right)\\
 & = & \left(y_l - y_l^0(\mathbf{b})\right)^2
\end{array}$$

since $\sum_{k\in U} b_k(y_k - y_k(\mathbf{b}))\mathbf{x}_k' = 0$. Then under the constraint (30), we get

$$\begin{array}{rlcl}
 & \left(y_l - y_l^0(\mathbf{b})\right)^2 - \gamma\frac{1}{(b_l+1)^2} & = & 0\\
\Leftrightarrow & \left(y_l - y_l^0(\mathbf{b})\right)^2 - \gamma\pi_l^2 & = & 0\\
\Leftrightarrow & \pi_l & = & \sqrt{\gamma}|y_l - y_l^*(\boldsymbol{\pi})|
\end{array}$$

where γ denotes a Lagrange multiplier. The result follows by application of constraint (11).

B Proof of Theorem 1

For any $t = 0, \ldots, T$, denote $\mathbf{b}^t = (b(\alpha_1^t), \ldots, b(\alpha_i^t), \ldots, b(\alpha_I^t))'$. Let $\mathbf{u} = (u_1, \ldots, u_i, \ldots, u_I)'$ be any $I \times 1$ vector, and

$$W_1(\mathbf{u}) = \frac{N}{N-q} \sum_{i=1}^{I} N_i u_i \sigma_i^2(\boldsymbol{\alpha}^{t-1}).$$

The minimization of $W_1(\mathbf{u})$ in $\mathbf{u}$, subject to

$$\sum_{i=1}^{I} \frac{N_i}{u_i + 1} = n \tag{32}$$

leads to $\mathbf{u} = \mathbf{b}^t$. Since $\mathbf{b}^{t-1}$ also satisfies equation (32), we have

$$W_1(\mathbf{b}^t) \leq W_1(\mathbf{b}^{t-1}) = V(\boldsymbol{\alpha}^{t-1}). \tag{33}$$

Now, let

$$W_2(\boldsymbol{\beta}) = \frac{N}{N-q} \sum_{i=1}^{I} b(\alpha_i^t) \sum_{k \in U_i} (y_k - \mathbf{x}_k' \boldsymbol{\beta})^2$$

where $\boldsymbol{\beta}$ denotes a $q \times 1$ vector. This is a standard fact that $W_2(\boldsymbol{\beta})$ is minimized by $\boldsymbol{\beta}^t = \left(\sum_{i=1}^{I} b(\alpha_i^t) \mathbf{A}_i \right)^{-1} \sum_{i=1}^{I} b(\alpha_i^t) \mathbf{c}_{1i}(y)$. Consequently, we obtain

$$W_2(\boldsymbol{\beta}^t) \leq W_2(\boldsymbol{\beta}^{t-1}) \tag{34}$$

where $\boldsymbol{\beta}^{t-1} = \left(\sum_{i=1}^{I} b(\alpha_i^{t-1}) \mathbf{A}_i \right)^{-1} \sum_{i=1}^{I} b(\alpha_i^{t-1}) \mathbf{c}_i(y)$. Since $W_2(\boldsymbol{\beta}^t) = V(\boldsymbol{\alpha}^t)$ and $W_2(\boldsymbol{\beta}^{t-1}) = W_1(\mathbf{b}^t)$, the result follows by a joint application of (33) and (34).

Références

Berger, Y. G., Munoz, J. F., Rancourt, E., 2009. Variance estimation of survey estimates calibrated on estimated control totals. an application to the extended regression estimator and the regression composite estimator. Computational Statistics and Data Analysis 53 (7), 2596 – 2604.

Bourdalle, G., Christine, M., Wilms, L., 2000. Echantillons maitre et emploi. Série Insee Méthodes 21, 139–173.

Chauvet, G., Tillé, Y., 2006. A fast algorithm for balanced sampling. Comput. Statist. 21 (1), 53–62.

Chauvet, G., Tillé, Y., 2007. Application of fast sas macros for balancing samples to the selection of addresses. CSBIGS 1, 173–182.

Chen, X.-H., Dempster, A. P., Liu, J. S., 1994. Weighted finite population sampling to maximize entropy. Biometrika 81 (3), 457–469.

Deville, J.-C., 2000. Note sur l'algorithme de Chen, Dempster et Liu. Tech. rep., CREST-ENSAI.

Deville, J.-C., Särndal, C.-E., 1992. Calibration estimators in survey sampling. J. Amer. Statist. Assoc. 87 (418), 376–382.

Deville, J.-C., Tillé, Y., 2004. Efficient balanced sampling : the cube method. Biometrika 91 (4), 893–912.

Deville, J.-C., Tillé, Y., 2005. Variance approximation under balanced sampling. J. Statist. Plann. Inference 128 (2), 569–591.

Hájek, J., 1964. Asymptotic theory of rejective sampling with varying probabilities from a finite population. Ann. Math. Statist. 35, 1491–1523.

Matei, A., Tillé, Y., 2005. Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. Journal of Official Statistics 21 (4), 543–570.

Matei, A., Tillé, Y., 2006. The r 'sampling' package. In : European Conference on Quality in Survey Statistics, Cardiff.

Tillé, Y., Favre, A.-C., 2005. Optimal allocation in balanced sampling. Statist. Probab. Lett. 74 (1), 31–37.